Spring 5-15-2017

Disordered Proteins: Connecting Sequences to Emergent Properties

Tyler Scott Harmon
Washington University in St. Louis

WASHINGTON UNIVERSITY IN ST. LOUIS

Department of Physics

Dissertation Examination Committee:
Rohit V. Pappu, Co-Chair
Ralf Wessel, Co-Chair
Greg Bowman
Anders E. Carlsson
Zohar Nussinov

Disordered Proteins: Connecting Sequences to Emergent Properties

by

Tyler S. Harmon

A dissertation presented to
The Graduate School
of Washington University in
partial fulfillment of the
requirements for the degree
of Doctor of Philosophy

May 2017

Saint Louis, Missouri

Table of Contents

Acknowledgments

There is no illusion in my mind that my career has been done in a vacuum, scientifically or otherwise. I have had the ability to gain guidance, insights, and moral support from countless people at every level of my career. I would like to highlight a few of the people who have influenced me the most.

My advisor, Rohit Pappu, has been an extraordinary guide into the world of being a researcher. He showed a remarkable patience and persistence in the early years of me being a graduate student where my ideas needed what can only be described as a translator to be understood by others. Much of my ability to effectively communicate complex ideas to people who approach science from a different perspective from myself is a result of Rohit's continual push for myself, and everyone in his lab, to be integrated into the scientific community through meeting with visiting professors, sending us to many conferences, and having active roles with many collaborators. Rohit has lead the lab as a prime example for what it means to be a scientist. More than any other influence in my life, Rohit's guidance has built the foundation of how I approach questions in science, build models of how the world works, and then attempt to tear them back down.

There are two groups of people in/from Rohit's lab that both deserve their own credit for their impact in my career. Albert Mao, Rahul Das, and Nick Lyle, all three of whom were nearing the end of their stays while I was getting started helped me transition from a graduate student taking classes to a graduate student that was doing research. Discussions with them help change my thinking from 'lets answer the question put in front of me' to 'what questions should we be asking'.

The second group in the lab is Kiersten Ruff and Alex Holehouse, both of whom joined the lab at roughly the same time as me. The countless discussions that we have had has helped build a picture of what is important in biophysics outside of the small pieces that I work on and helped me integrate the field into a more coherent understanding. They have also helped me think about many problems and helped me develop solutions that I could not have done on my own. I'd also like to thanks Ammon Posey who, in addition to thoughtful discussions, has taken our crazy theories and turned them into tangible experiments.

I also want to acknowledge that outside Rohit's lab there has been a supportive community in the Washington University physics department that is dedicated to the success of its students. This culture has led to both many supportive graduate students as well as professors. Specifically I would like to thank the rest of my thesis committee, Anders Carlsson, Ralf Wessel, Zohar Nussinov, and Greg Bowman. The four of you have been a large part of the support from the faculty and for many discussions and lectures on topics that I otherwise would not have found an interest in.

Finally, I would like to thank my family for all their support over the years. Without their moral support, I never would have gotten this far.

Tyler Harmon

Washington University

May 2017

Chapter 1

Introduction

1.1 Overview

In 1943 Erwin Schrödinger gave a series of lectures at Trinity College in Dublin. These lectures were published in 1944 as a monograph entitled "What is life?" [1]. Schrödinger was not looking to answer metaphysical questions. Instead, he wished to explain how "…*the events in space and time which take place within the spatial boundary of a living organism can be accounted for by physics and chemistry*?" Although Schrödinger got many details wrong, the structure of his thinking is informative. He was interested in uncovering a unifying physical picture of living systems. He introduced key features of living systems: these included stochasticity at the smallest length scales. "Order from disorder" was the theme as was epitomized, for Schrödinger, in nature solving numerous problems using diffusion. Inheritance was the next theme and he built on what was known about the genetic basis of life and the concept of the so-called "hereditary molecule". Principally, Schrödinger envisaged molecules as the carriers and propagators of information that are needed to support the essentials of living systems namely, *replication*, *development*, and *growth*. He recognized the importance of spatial and temporal organization of molecules into nested hierarchies. Schrödinger reasoned that molecules have to be the agents that organize into information receiving, processing, and transduction units and these units would have to be out of equilibrium, operating under the influence of sources and sinks.

We have come a long way from Schrödinger's conceptions of what makes up living systems. The building blocks of life are indeed molecules and they include DNA, RNA, proteins, lipids, sugars, and other small molecules. Uncovering the synergy between encoded information and environmental cues as determinants of driving forces for and rates of self-assembly is a daunting enterprise. Indeed, the past seven decades of research at the multi-way interface of molecular biology, chemistry, and physics, captured perfectly in the field of molecular biophysics, has helped us understand many of the building blocks of life in excruciating detail. The persistent motto has been that we should be able to work out all the details of interactions and self-assembly at the molecular level and put the parts back together to reconstruct a living system. At a fundamental level this is not an erroneous axiom or precept. For instance, if I were to take apart a car down to every single part and hand these parts to a friend, then, in theory, that friend should have all the ingredients necessary for putting the car back together. My friend would also know that if she were to succeed in putting the parts together, then the end result should be an entity that looks like a car, which should run when fuel is provided. From a didactic sense, all the information that is necessary to put the parts together and reconstruct the automobile powered by fuel injection and internal combustion are essentially there, once my friend figures out how the parts fit together. Of course, a blueprint of the assembly would be enormously helpful and would go a long way toward expediting the process of putting the car back together. However, my friend is rather smart, and with unlimited time at her disposal, she should be able to reconstruct the car from its parts.

In the preceding didactic example, knowledge of the end-state certainly helps, but also crucial is the knowledge that the parts are not going to undergo non-trivial changes as a function of time. For instance, the parts are not going to come together to make new parts. In response to

some external stimulus or cue, the parts are not going to switch on “genetic components”, turn on their “expression”, and “translate” the information in the genetic components into a new set of parts. Conventional physical systems that obey the laws of equilibrium thermodynamics are not known for emergent properties, for adaptation, for responsiveness over a spectrum of spatial and temporal scales, for self-replication, reproduction, and of central interest to this thesis work – morphogenesis on different scales. Even the simplest unicellular organisms have incredible spatial organization and dynamics that characterize their entire life cycles [2]. The molecular building blocks are compartmentalized into information storage and processing depots [3]. These depots are built around molecules large and small. They come together in response to cues and the reactions that are catalyzed, sequestered and managed in depots or so-called organelles determine a variety of cellular decisions and fates [4]. They contribute to emergent properties that beyond the cellular level and control tissue-level organization and organ development [5].

Ultimately, morphogenesis, the achievement of distinct morphologies as a function of time, through the collective interactions amongst the molecular building blocks of life, may be viewed as the stepping stone toward an understanding of how living systems are born, how they adapt, how they reproduce, and how they manage their own demise through resource management and aging programs. Not surprisingly, the new burning question for physics is actually a rather old one, first posed formally by Schrödinger, but one that occupied the interest of many influential thinkers before and after Schrödinger.

1.2 Reducing Life to its Parts

Molecular biophysics and molecular genetics, when put together, hold the promise of getting us closer to answering the question posed by Schrödinger. Geneticists are extremely good

at detective work: They observe an organism of interest or a cellular process of relevance and through the marvels of molecular biology, they are capable of uncovering the genes that, when deleted or altered, show a direct connection to the behavior of interest. These genotype-to-phenotype relationships are the bedrock of molecular biology and genetics. They provide the starting point for almost every investigation that opens the door to drawing direct connections between molecular building blocks and the property or behavior of interest.

Sometimes, it transpires that the information encoded in genes proves insufficient to connect to the phenotype of interest. One has to take so-called epigenetic factors into account. In their simplest conception, epigenetics refers to the rewriting of genes by environmental factors such as the oxidative damage of DNA, which increases DNA methylation. These details aside, the rules of the game are invariably the same: Find the gene, the epigene, or cluster of genes / epigenes and connect these to phenotype. With the genotype-to-phenotype connections in hand, we segue into the world of molecular biophysics, which is all about uncovering the physical principles that connect the genotype-to-phenotype. What starts off as a goal that should, at least in some far-removed sense, be connected to answering the central question of interest, often becomes a descent down a gopher's burrow. The building blocks of life are incredibly complex in terms of the amount of information that they potentially encode. This is manifest in the achievement of specific structures, the interactions among these structures, and the dynamics of these structures. An entire research career can be dedicated to the study of one molecule and its complex array of interactions be it a protein, an RNA, or genomic DNA. The further down the gopher's burrow we descend, the farther removed we become from the driving questions of morphogenesis and the physics that gives rise to emergent properties.

1.3 A Way Out of Reductionist Thinking

While biology and to a large extent chemistry have become increasingly more reductionist, many of the advances in physics have focused on the description of emergent phenomena as well as adaptive and responsive processes. Nowhere is this more evident than the continually evolving fields of phase transitions, the description of critical phenomena, and nonlinear dynamics of non-equilibrium systems [6]. Each decade since the 1960s has seen significant leaps being made in these areas and the physics of soft-matter has matured and spawned the physics of active matter [7], which is bringing us close to figuring out living systems self-organize, adapt, and replicate in response to cues and "know / learn" to protect themselves to unpredictable stresses. For a while, the impact of phase changes and the underlying physical principles were not fully appreciated or even deemed to be relevant for describing biological systems. However, as the genetic programs were uncovered, it became clear that there needs to be a framework for describing collective phenomena and synergies among various units across multiple length and timescales to connect information encoded in genes to morphogenesis and higher order development [8].

1.4 Two Paradigm Shifts Anchor Physics to Biology

Structural biology has been the bedrock of the interface between physics and biology. Among the major driving problems was the pioneering observation made by Anfinsen that proteins, the molecules of life, could spontaneously fold into their functional, albeit irregular three-dimensional structures [9] . Since the pioneering experiments of Anfinsen, the reversibility of protein folding has been demonstrated for numerous proteins. In parallel with protein folding, there was another revolution afoot in structural biology. Proteins could be crystallized and

shining x-rays on these crystals with irregularly shaped molecules occupying the asymmetric units revealed protein structures to atomic level detail. Thus came about the sequence-structure-function paradigm. A protein must fold into a well-defined three-dimensional structure in order to be able to perform its biological function. The information encoded in the primary sequence, in synergy with appropriate solution conditions, is sufficient to enable folding into the structure that is biologically active. Therefore, the central questions became: How do proteins fold? How many distinct folds are available to proteins? Why do proteins that are vastly different in sequence converge upon similar folds? And how do we understand a protein's function from its structure? To these questions, there was the added realization that proteins are not rigid entities. They undergo spontaneous thermal fluctuations and these motions appear to be biologically relevant as well. Hence, the sequence-structure-function paradigm could be generalized to be the pursuit of sequence-structure-dynamics-function relationships [10]. If we knew all the structures for all the sequences and all of the relevant motions encoded by sequence-structure relationships, we could, most certainly work out the molecular functions of proteins and build up to answering questions about higher order processes.

However, the revolution in genome sequencing revealed a rather surprising set of findings: First, there simply aren't enough proteins encoded by our genomes – at least at first glance – to account for all the functions that are attributable to proteins. Second, higher order genomes seem to be enriched in coding regions that yield proteins that are intrinsically disordered [11]. As autonomous units these proteins do not adopt a singular, well-defined three-dimensional structure under physiologically relevant conditions. Closer inspection revealed that these so-called intrinsically disordered proteins (IDPs) were everywhere in that they were involved a whole range of functions, crucially relevant to the regulation of higher order

processes that are central to connecting genetic information to morphogenetic processes through emergent properties [12].

Some IDPs are deferred folders; in the presence of an appropriate ligand, they fold into well-defined three-dimensional structure. The wishful thinking was that all IDPs undergo coupled-folding and binding in order to function [13], but these appear to be a sizable minority of the intrinsically disordered proteome. A larger fraction of these proteins seem to serve as linkers, bristles, actuators, scaffolds, adaptors, loci of facile molecular recognition, and targets of posttranslational modifications that serve as interactions hubs thus coordinating a range of crucial cellular functions and processes [14, 15]. IDPs therefore defy the conventional sequence-structure-dynamics-function paradigm. There clearly is much more in store at the molecular level and indeed there appears the possibility of encoding the ability for nonlinear transfer of information from genes to processes through multiplicity of interactions (we return to the theme of order from disorder) and IDPs seem to be suitable candidates for transducing information.

A second challenge to the importance of well-defined structures being the molecular and supramolecular building blocks of life came from the recognition that spatial organization within cells also defies standard ideas of linear transformations from building blocks to organelles. Cells are rich with sub-cellular compartments. Many of these compartments that are routinely discussed in biology text books are membrane bound and include objects like the golgi, endosomes, mitochondria and the like. Clearly, compartmentalization is essential to sequester molecules and gain efficiencies of reactions and lipids, with their surfactant-like character seem to be ideal candidates for forming vesicular compartments.

However, the cell is actually replete with numerous organelles or substructures that do not have a vesting membrane. These, so-called, membraneless organelles are composed either

entirely of proteins or proteins and nucleic acids. They include the nucleolus, nuclear speckles, the centrosome, the pericentriolar material, signalosomes, P-bodies, P-granules, RNA stress granules, Balbiani bodies, Cajal bodies, nuage bodies, and several other micron- and sub-micron-scale objects [5, 16-20]. Over the past several decades, biologists have shown that these bodies exist and that they are essential for a range of functions that control crucial cellular processes that give rise to emergent properties on higher order length scales. However, even though the identities of molecular components of many of the bodies were coming into focus, there were three major unanswered questions: How do these bodies form? Why do distinct bodies accumulate certain protein and nucleic acid molecules and exclude others? And importantly, how do these bodies contribute to cellular functions including cellular responses as well as homeostasis?

In 2009 there came a major breakthrough. Brangwynne and Hyman were studying the developing embryo in *C .elegans* and noticed an asymmetric partitioning of molecules within the developing embryo [16]. The body of interest, the so-called P granule was crucial for this asymmetric accumulation and importantly, this micron-scale protein-RNA body seemed to have all the material properties of a liquid droplet. It was spherical; it fused with other droplets; it flowed and the molecular components exchanged with the surroundings on times scales that defied description in terms of a rigid scaffold. Since this epic discovery, there have been several major findings, all pointing to the panoply of micron-scale liquid-like bodies within cells. Strikingly, these bodies seem to form as the result of phase transitions, whereby key proteins drive a demixing transition from the cytoplasm or nucleoplasm leading to the formation of a dense liquid droplet rich in macromolecular components that is in equilibrium with a dilute phase that is deficient in macromolecules. The conserved order parameter is macromolecular

concentration and there is a finite interfacial tension between the two phases. Phase separation is reversible and appears to be under biological control in that the expression of proteins that drive phase separation is highly regulated as are processes that can dilute the droplets and dissolve them when they are no longer needed. Importantly, many of the proteins that drive phase separation, if not all, involve large stretches of intrinsically disordered regions.

1.5 Shifts in Paradigms Provide the Basis for Integrating Molecular and Meso Scales

And now we come to the two synergistic foci of this thesis IDPs and the phase transitions that they drive, because these help us ascend above reductionist investigations and think about higher order processes that drive morphogenesis and development. My thesis work covers a spectrum of problems – all involving IDPs and their interactions: I have contributed advances to the problem of coupled folding and binding by showing that one can design the extent of pre-foldedness in an IDP and assess the extent to which pre-organization helps drive coupled folding and binding reactions. I have contributed key insights to the problem of discerning the types of conformations that intrinsically disordered regions (IDRs) can adopt when tethered to ordered domains. I have built on these insights and developed a computational framework to predict phase diagrams for linear, multivalent proteins that drive phase separation. My focus in this work has been on modulating the convolution between sol-gel transitions and liquid-liquid phase separation by disordered linkers. The methods developed here have enabled an explanation for the sub-structures observed in nucleoli thus showing that layered organelles do not have to be solids – they can be coexisting liquids [21]. I further showed that the synergy between IDPs and ordered domains in so-called partially ordered polymers that have been engineered to undergo

thermally responsive phase transitions can also demonstrate hysteresis in their phase transitions. This hysteresis is controllable and I uncovered the basis for this control, thus opened the door to designing memory effects through molecular switches. Finally, I also discovered a surprising molecular feature of IDP sequences: The information written into a sequence is not static. Charge renormalization (the alteration of the total charge on the molecule by solution ions) and charge regulation (alteration of specific charges by proton release or uptake) can have a profound effect on disorder to order transitions of putative IDPs. Interestingly, many of the molecules that serve as scaffolds for driving phase separation and enabling cell migration use sequences that are likely to undergo charge regulation. Taken together, the totality of my thesis work provides a foundation for a variety of investigations that are based on the synergy between IDPs and phase behavior. This holds the promise of enabling the discovery of physical principles that underlie morphogenesis, development, adaptation, and replication.

1.6 Information is Key

Information is the driver of life because it can be created, stored, processed, and transferred. Importantly, thanks to the pioneering work of Claude Shannon [22] and Edwin Jaynes [23], we know that information can also be quantified and this ability allows us to connect information to thermodynamic driving forces, such as entropy [23], irreversibility in dissipative systems [24], and dynamical quantities such as caliber [25]. Therefore, given knowledge of the building blocks of life, their interactions, conversions, and transformations, one can start to build a model for information flow around what we learn from the basic interactions amongst molecular building blocks and their underlying dynamics.

Life arises from the propagation of information and information that is not propagated eventually dies out in favor of information that does propagate. DNA, RNA, and proteins are nothing more than a means to an end in the propagation of information. Successful propagation of this information requires coordination of *information generation, storage, processing, and transfer* in cellular systems.

At the length scale of molecules, a protein can be thought of as a functional bit that the cell uses for information management. In this case, the information unit is generated through the combined processes of transcription and translation. Once the information unit has been generated, the protein can interact with specific binding partners and this is an information transfer process. By interacting within cells, proteins serve as molecular conduits of information within the cell. A lot of work has gone into understanding the structure of proteins with the prospect that knowing the structure of the protein will give a direct understanding of how the protein works. This has been the conventional wisdom in biological science and has been termed the "structure-function" paradigm. This idea is especially easy for human digestion because our tools are all constructed out of solid materials. The structure of a wrench and bolt can help explain much of how we use wrenches.

The idea that a protein folds into a single shape has been shown to miss many important features for proteins [26]. Specifically, ensembles of conformation provide better descriptions of proteins, where each conformation has a statistical weight governed by the Boltzmann distribution. Ensembles and their statistical weights can be understood through the prism of information theory, because the microstates within an ensemble define the communication channels for information transfer and the weights associated with binding competent microstates refer to the channel capacities. Additionally, when the magnitude and type of fluctuations on

two different surfaces of a protein are coupled together, then binding to one of these regions changes the properties of the other. It has been shown for several proteins that these types of coupled sites can share information across entire sides of proteins. This complicates the understanding of how proteins work and integrate information because in order to understand protein function from a molecular perspective and uncover its role in information transfer within the cell, one must uncover the properties of the ensembles that are encoded by the energy landscape of a protein [26].

The idea that the ensemble can be more important than the dominant conformation extends much further into the class of intrinsically disordered proteins [27]. These proteins have no tertiary structure and weak secondary structure if any at all in large stretches of residues. The disorder encoded by the amino acid sequence leads to significant degrees of heterogeneity in the ensembles [28]. A persistent question in the field is, how does being disordered affect the ability and ways that information can be transferred and integrated by cells? The conventional approach to uncover answers to this question has been to study the binding mechanism. This involves the deployment of investigations that quantify the driving forces for coupled folding and binding and quantifying the fluxes through paths that go from heterogeneous monomeric ensemble to a homogeneous bound ensemble. Coupled folding binding is often thought of from the vantage point of two limiting kinetic scenarios, viz., conformational selection or induced fit [29]. In the former, the IDP has to first fold before it docks to its target and in the latter scenario, a liquid-like interface defined by correlation functions precedes the acquisition of the folded complex defined by specific stable contacts. Of course, these are limiting scenarios, when in fact the acquisition of structure could be defined by a continuum of possibilities.

How might the mechanisms of coupled folding and binding contribute to the regulation of cellular processes? One possible explanation might be that the partial ordered states play an important role in the transition between the interactions of two competing ligands. Having a semi-stable state that is half ordered and half disordered, cells can create an intermediate whereby both competing ligands are partially bound. This has the potential to speed up the equilibration under new conditions because there would no longer be the need for an absolute maximum re-equilibration rate of the off rate for the cognate ligands. If proteins are using folded domains for the same competitive binding signal, it is easy to imagine a case where the original ligand must be completely dissociated before the competitor is able to attempt to bind. The presence of an intermediate where both are bound is a unique feature that folded proteins cannot easily achieve.

With the current arsenal of tools at the disposal of researchers, the hypothesis that disorder fosters the exchange from one ligand bound state to another remains elusive for investigation. Chapter two of this thesis describes a methodology that should enable systematic investigations of coupled folding and binding by enabling the systematic titration of foldedness within an IDP by maintaining the overall amino acid composition and altering the predisposition toward a chosen structure while satisfying prescribed constraints. In our case study, we designed sequences where the intrinsic helicity is stabilized/destabilized in the two halves of a protein that has been used as a model system for coupled folding and binding. We currently have ongoing collaborations where we are measuring the thermodynamics and kinetics of these designed ligands binding with their partner in order to detect changes in the binding pathway. We hope that these designed sequences and methodology will be used to answer how disorder affects the kinetics of binding and unbinding. More importantly, we think these designed proteins will help

shed light on why these proteins are disordered through testing if there is a functional change in cells when the degree of heterogeneity is modulated.

At a higher level, cells have to transfer information through moving components with respect to the cell or exert forces on the cytoskeleton. A few examples of such processes are moving cargo when diffusion is too slow, creating contraction forces across the cell, and remodeling the cytoskeleton. Motor proteins known as myosins are central to many of the processes that involve the generation of forces. A common theme of these proteins is the presence of a long rod like segment that connects two folded domains [30-34]. These domains bind to the cytoskeleton on one side and either another segment of cytoskeleton or a component that has been designated by the cell as needed transport. Phosphorylation and dephosphorylation change the preferred equilibrium angle of this long segment with respect to the two folded domains. Cells use this to make the rod like segment swing back and forth like a long leg. The rod like segment acts as a lever arm in exerting force. Current models for how these movements work requires that this rod segment be stiff. If it has a persistence length on the length scale of the rod then the models break down into a highly inefficient method for force inducement because the change in angle does not cause much linear movement at the other end of the rod. As such, this rod needs to be closer to a molecular 2 ×4 as opposed to a pool-noodle.

One archetype that is used in these motors for a molecular 2 ×4 is a repeating pattern of Glu and Lys residues namely $[(Glu)_4\text{-}(Lys)_4]_n$ where n is the number of repeats [30-34]. This sequence makes a long, rigid alpha-helical rod. Chapter 3 is devoted to understanding how this sequence forms a stiff rod. The coil-to-rod transition involves significant charge regulation, whereby the preferential protonation of internal Glu residues leads to the formation of a stable alpha helix whereas the fully neutral sequence adopts a heterogeneous coil-like ensemble of

conformations. This work sets the stage for understanding the full extent of sequence-encoded charge regulation that gives rise to disorder to order or order to disorder transitions and their roles in regulating key cellular processes.

Membraneless organelles provide intracellular spatial organization and the compartmentalization that is needed for functional robustness [17, 18, 35-39]. In membrane-bound organelles, the membrane provides a physical barrier to differentiate between inside and outside the organelle. Organelle specific channels, pumps, and sensors decorate these membranes and these are used to transmit and integrate information from the cell to control what the organelle is doing at any given time. This is an appealing method for organizing functions that can benefit from being compartmentalized because it has parallels to how we build our factories. Different processes are carried out in different areas and walls and conveyer belts are used to move components to and from areas that benefit from being manufactured in different areas.

Over the past decade, it has emerged that there is a large class of organelles that form in cells that have no membrane-bound walls. They have clearly distinct boundaries that separate the inside from outside of the organelle and they are effective in controlling what goes into and out of them. These organelles are composed primarily of proteins and RNA that has phase separated into polymer based liquid-like droplets. A basic theme of these proteins is their apparent multivalency where they can simultaneously interact with multiple proteins at a time. As such they have been conveniently termed scaffold proteins. The sequence-encoded preference for multivalent interactions provides the driving force for phase separation. This is believed to be the dominant mechanism for the formation of these organelles.

Membraneless organelles have a significant advantage over membrane bound organelles in that changes in the cell can significantly change in the interaction strength between proteins and or RNA. This means that cells can regenerate and dissolve membraneless organelles, which are also referred to as biomolecular condensates, by changing protein or RNA levels, by posttranslational modifications to proteins, by posttranscriptional processing of RNA, and by regulating the half-lives of the proteins and RNA molecules. When a cell feels the need to respond in a different way to the environment or to internal conditions, it can "call upon" multivalent protein and RNA molecules to phase separate and construct a new organelle or dissolve an existing organelle for a specific function [19, 40]. Because phase separation has a well-defined critical concentration, the cells have a convenient on / off switch for functionality.

In a consistency with the theme of phase separating from the rest of the cell, many of the reasons why the cell constructs / deconstructs these organelles are associated with events that scale across the entire cell. Some examples are cellular stresses and cell division [41, 42]. Understanding what types of properties are important for phase separation can give insights into the types of mechanisms that we should be looking for in how cells prepare and execute these responses. As such, chapter 4 is focused on predicting the role of disordered regions that tether protein-protein interaction domains in linear multivalent systems. Additionally, the work points to an alternative behavior that cells could be using. While phase separation has obvious implications for bringing components that work together into a small volume or storing components in an inert volume for protection or later use, these disordered regions can also push cells into the less obvious response of forming a physical gel. This is a state where proteins have crossed a critical point where they are fully networked across the cell [43]. This would mean that from a given protein you would be walk across connected proteins to any other spot in the

cell. This affects the kinetics of the cell more distinctly than the thermodynamics so if this type of response is real it is most likely associated with a temporary halting of processes in the cell under extreme stress such as starvation.

If we have cells phase separating proteins from the rest of the cytoplasm and nucleus, it raises the question of if there is internal organization inside a membraneless organelle. Having an organization inside a droplet could be a useful way to integrate another layer of information from the cell. Chapter 5 is devoted to collaboration with the Brangwynne lab where they observed that the nucleolus, a structure in the nucleus, is a liquid-like droplet with liquid-like droplets inside. We propose that this type of architecture holds significant information for the cell about constructing ribosomes, the protein factories of the cell. The initial RNA building blocks to ribosomes are transcribed in the inner droplets and then diffuse ultimately out of the nucleolus. This has them diffusing from the innermost layer, through layers of different chemical environments. This gives the cell the ability to control the order of operations on post-transcriptional modifications.

It is known in the ribosome field that assembly is more complicated than mixing the proteins and RNAs together in a test tube under cellular like salt and protein concentrations. There are large kinetic barriers to their formation that experimentalists can overcome through heating and cooling regiments. It is also known that many posttranscriptional modifications are localized to the core region or exclusively in the shell region. We propose that controlling the order of operations through this spatial organization inside the droplets gives cells the needed control to assemble the ribosomes without large barriers. By introducing post-transcriptional modifications part way through the assembly, we think that cells can circumvent the kinetic traps that experimentalists observe.

Additionally, by localizing the components into these layers gives cells an added mechanism for transferring and integrating signals in the cell. Dissolving the inner droplet into the outer droplet could act as an effective response mechanism under certain conditions. It is known that the nucleolus as a whole dissolves under cell division but it is not known if this spatial organization holds intact under different stresses or cellular responses.

The previous sections have largely revolved around mechanisms that exploit thermodynamic aspects of phase transitions. The dynamics of phase separation, controlled by energetic, diffusive, and topological barriers will also determine the phase behavior of protein-based polymers. Tropoelastins are examples of proteins that encode viscoelastic properties upon coacervation and chemical crosslinking. The Chilkoti lab has attempted to synthesize minimalist mimics of tropoelastins using polymers of pentapeptide repeats of elastin-like polypeptides (ELPs) interspersed by alpha-helical polyalanine domains. These molecules are thermally responsive protein based block co-polymers that show tunable hysteresis in their thermal transitions. The lower critical solution temperatures (LCSTs) measured along the heating and cooling arms are tunable and can be non-overlapping. The number and type of alanine-rich blocks determine the extent of hysteresis. We have developed a phenomenological model that reproduces the experimentally observed tunable hysteresis for block copolymeric sequences. This requires an imbalance between the strengths of homotypic interactions between alanine-rich regions and ELP repeats. Additionally, the ELPs and alanine-rich regions have to be immiscible with one another. These features engender micro-phase separation whereby the block copolymers form spherical clusters comprising of alanine-rich cores and ELP coronas. Upon raising the temperature above the LCST, the clusters are drawn to one another by favorable interactions among ELPs and these clusters further network via domain swapping of alanine-rich

regions. Lowering the temperature below the LCST leads to dispersion of the clusters by weakening of homotypic interactions between ELPs. However, the domain swapped states persist and this maintains the physical networking of clusters thus giving rise to hysteresis. Domain swapping and the persistence of this state below the LCST are governed by the energy gap between the homotypic interactions of ELPs vis-à-vis alanine-rich regions. Our findings, presented in chapter 6, have direct bearing on the *de novo* design of responsive materials based on IDPs, where tunable hysteresis can be used to encode memory effects, and for understanding the complexities of sequence-encoded phase behavior of archetypal low complexity disordered proteins.

1.7 Concluding Words for the Prologue

Before launching into the chapters and their intricate details, I would like to highlight the fact that much of the work reported here involved an integrated effort, often motivated by experimental observations and developments that came along during the course of this dissertation. Indeed, a quick glance at my thesis proposal will reveal that I veered away from what I proposed to do for my thesis work because I followed my interests and went where the questions seemed most intriguing, current, and impactful. I have insisted on seeking problems that are foundational and if solved could help lay the foundations for the larger-scale problems that truly interest me. Such an endeavor is, by necessity, collaborative and integrative. Therefore, each chapter represents a synthesis of my contributions to a particular problem, but it also encompasses the contributions of my colleagues within the lab and beyond. This integration is essential to ensure that this narrative is coherent and complete. To highlight the synthesis of multipronged efforts and multiple contributions, I shall use the term “we” to refer to all of the work being reported here, even in cases where the work happened to be exclusively mine. This

inclusiveness ensures that the valuable contributions made by my colleagues to providing shape to this thesis.

1.8 References

1. Schrödinger, E., *What is Life? The Physical Aspect of the Living Cell*. Based on lectures delivered under the auspices of the Dublin Institute for Advanced Studies at Trinity College, Dublin, in February 1943. 1944: Cambridge University Press. 90-165.
2. Kerfeld, C.A., et al., *Protein Structures Forming the Shell of Primitive Bacterial Organelles.* Science, 2005. **309**(5736): p. 936-938.
3. Diekmann, Y. and J.B. Pereira-Leal, *Evolution of intracellular compartmentalization.* Biochem J, 2013. **449**(2): p. 319-31.
4. Ramaswami, M., J.P. Taylor, and R. Parker, *Altered Ribostasis: RNA-Protein Granules in Degenerative Disorders.* Cell. **154**(4): p. 727-736.
5. Hyman, A.A. and C.P. Brangwynne, *Beyond stereospecificity: liquids and mesoscale organization of cytoplasm.* Dev Cell, 2011. **21**(1): p. 14-6.
6. Yeomans, J.M., *Statistical mechanics of phase transitions*. 1992: Clarendon Press.
7. Ramaswamy, S., *The Mechanics and Statistics of Active Matter.* Annual Review of Condensed Matter Physics, 2010. **1**(1): p. 323-345.
8. Cates, M.E. and J. Tailleur, *Motility-Induced Phase Separation.* Annual Review of Condensed Matter Physics, 2015. **6**(1): p. 219-244.
9. Anfinsen, C.B. and H.A. Scheraga, *Experimental and theoretical aspects of protein folding.* Adv Protein Chem, 1975. **29**: p. 205-300.

10. Campbell, E., et al., *The role of protein dynamics in the evolution of new enzyme function.* Nat Chem Biol, 2016. **12**(11): p. 944-950.

11. Wright, P.E. and H.J. Dyson, *Intrinsically unstructured proteins: Re-assessing the protein structure-function paradigm.* Journal of Molecular Biology, 1999. **293**(2): p. 321-331.

12. Dunker, A.K., et al., *Intrinsically disordered protein.* J Mol Graph Model, 2001. **19**(1): p. 26-59.

13. Oldfield, C.J., et al., *Coupled folding and binding with alpha-helix-forming molecular recognition elements.* Biochemistry, 2005. **44**(37): p. 12454-12470.

14. Romero, P., Z. Obradovic, and A.K. Dunker, *Natively disordered proteins: functions and predictions.* Appl Bioinformatics, 2004. **3**(2-3): p. 105-13.

15. Dyson, H.J. and P.E. Wright, *Intrinsically unstructured proteins and their functions.* Nature Reviews in Molecular Cell Biology, 2005. **6**: p. 197-208.

16. Brangwynne, C.P., et al., *Germline P granules are liquid droplets that localize by controlled dissolution/condensation.* Science, 2009. **324**(5935): p. 1729-32.

17. Lee, C.F., et al., *Spatial organization of the cell cytoplasm by position-dependent phase separation.* Phys Rev Lett, 2013. **111**(8): p. 088101.

18. Hyman, A.A., C.A. Weber, and F. Julicher, *Liquid-liquid phase separation in biology.* Annu Rev Cell Dev Biol, 2014. **30**: p. 39-58.

19. Banani, S.F., et al., *Biomolecular condensates: organizers of cellular biochemistry.* Nat Rev Mol Cell Biol, 2017.

20. Brangwynne, C.P., *Phase transitions and size scaling of membrane-less organelles.* J Cell Biol, 2013. **203**(6): p. 875-81.

21. Feric, M., et al., *Coexisting Liquid Phases Underlie Nucleolar Subcompartments.* Cell, 2016. **165**(7): p. 1686-97.

22. Shannon, C.E., *A mathematical theory of communication.* The Bell System Technical Journal, 1948. **27**(3): p. 379-423.

23. Jaynes, E.T., *Information Theory and Statistical Mechanics.* Physical Review, 1957. **106**(4): p. 620-630.

24. Jaynes, E.T., *Information Theory and Statistical Mechanics. II.* Physical Review, 1957. **108**(2): p. 171-190.

25. Hazoglou, M.J., et al., *Communication: Maximum caliber is a general variational principle for nonequilibrium statistical mechanics.* J Chem Phys, 2015. **143**(5): p. 051104.

26. Hilser, V.J., *An Ensemble View of Allostery.* Science, 2010. **327**(5966): p. 653-654.

27. Mao, A.H., N. Lyle, and R.V. Pappu, *Describing sequence-ensemble relationships for intrinsically disordered proteins.* Biochem J, 2013. **449**(2): p. 307-18.

28. Lyle, N., R.K. Das, and R.V. Pappu, *A quantitative measure for protein conformational heterogeneity.* J Chem Phys, 2013. **139**(12): p. 121907.

29. van der Lee, R., et al., *Classification of intrinsically disordered regions and proteins.* Chem Rev, 2014. **114**(13): p. 6589-631.

30. Knight, P.J., et al., *The predicted coiled-coil domain of myosin 10 forms a novel elongated domain that lengthens the head.* J Biol Chem, 2005. **280**(41): p. 34702-8.

31. Baboolal, T.G., et al., *The SAH domain extends the functional length of the myosin lever.* Proc Natl Acad Sci U S A, 2009. **106**(52): p. 22193-8.

32. Wolny, M., et al., *Stable single alpha-helices are constant force springs in proteins.* J Biol Chem, 2014. **289**(40): p. 27825-35.

33. Samejima, K., et al., *The Inner Centromere Protein (INCENP) Coil Is a Single alpha-Helix (SAH) Domain That Binds Directly to Microtubules and Is Important for Chromosome Passenger Complex (CPC) Localization and Function in Mitosis.* J Biol Chem, 2015. **290**(35): p. 21460-72.

34. Wolny, M., et al., *Characterization of long and stable de novo single alpha-helix domains provides novel insight into their stability.* Sci Rep, 2017. **7**: p. 44341.

35. Bergeron-Sandoval, L.P., N. Safaee, and S.W. Michnick, *Mechanisms and Consequences of Macromolecular Phase Separation.* Cell, 2016. **165**(5): p. 1067-79.

36. Mitrea, D.M. and R.W. Kriwacki, *Phase separation in biology; functional organization of a higher order.* Cell Commun Signal, 2016. **14**: p. 1.

37. Shin, Y., et al., *Spatiotemporal Control of Intracellular Phase Transitions Using Light-Activated optoDroplets.* Cell, 2017. **168**(1-2): p. 159-171.e14.

38. Zhu, L. and C.P. Brangwynne, *Nuclear bodies: the emerging biophysics of nucleoplasmic phases.* Curr Opin Cell Biol, 2015. **34**: p. 23-30.

39. Brangwynne, C.P., P. Tompa, and R.V. Pappu, *Polymer physics of intracellular phase transitions.* Nat Phys, 2015. **11**(11): p. 899-904.

40. Su, X., et al., *Phase separation of signaling molecules promotes T cell receptor signal transduction.* Science, 2016. **352**(6285): p. 595-9.

41. Kroschwald, S., et al., *Promiscuous interactions and protein disaggregases determine the material state of stress-inducible RNP granules.* Elife, 2015. **4**: p. e06807.

42. Munder, M.C., et al., *A pH-driven transition of the cytoplasm from a fluid- to a solid-like state promotes entry into dormancy.* Elife, 2016. **5**.

43. Parry, B.R., et al., *The bacterial cytoplasm has glass-like properties and is fluidized by metabolic activity.* Cell, 2014. **156**(1-2): p. 183-94.

Chapter 2

GADIS: Algorithm for Designing Sequences to Achieve Target Secondary Structure Profiles of Intrinsically Disordered Proteins

This chapter is adapted from an article[1] published in Protein Engineering Design and Selection. Michael Crabtree, Sarah Shammas, Ammon Posey and Jane Clarke designed and conducted the experiments. Tyler S. Harmon and Rohit V. Pappu developed algorithm framework. Tyler S. Harmon performed and analyzed the simulations.

2.1 Introduction

Many macromolecular complexes involve proteins or regions that are intrinsically disordered in their unbound forms [2-6]. Intrinsically disordered proteins / regions (IDPs / IDRs) are distinct from autonomously folded domains. The amino acid sequences of IDPs encode an intrinsic preference for conformational heterogeneity, which means that they do not fold into specific three-dimensional structures as autonomous units [7]. Many IDPs are involved in molecular recognition [8] and one mode of recognition involves coupled folding and binding [3, 9, 10]. Here we focus on a specific archetype, namely binary complexes where IDPs fold when they are bound to pre-folded protein partners.

A majority of IDPs that undergo coupled folding and binding tend to adopt α-helical structures in their bound complexes. Interestingly, many of these IDPs have quantifiable intrinsic helicities in their unbound forms [8, 11-14]. Recently, Borcherds et al. [15] showed that point

mutations could be engineered into the intrinsically disordered N-terminal domain of the tumor suppressor p53 to enhance its intrinsic helicity. This proline-to-alanine substitution leads to an increase in the affinity of p53 for Mdm2. Of course, a particular value for the dissociation constant (K_D) can accommodate a range of mechanisms for coupled folding and binding [16]. This feature is highlighted in kinetics experiments that have measured the rates of association of the intrinsically disordered BH3-PUMA (referred to hereafter as PUMA) peptide to the pre-folded Mcl-1 [17-19] and other systems [20]. Systematic proline and alanine scanning of PUMA was used to assess the contributions of helicity in unbound PUMA on the mechanisms of coupled folding and binding [18, 19]. Proline and alanine scanning do not significantly alter the association rates. However, the rates of dissociation (k_{off}) of PUMA from Mcl-1 show significant changes upon proline- or alanine-scanning mutations to the PUMA sequence.

An intriguing hypothesis is that the amino acid composition of an IDP is the main determinant of k_{on} whereas the degree of intrinsic helicity regulates k_{off} thus leading to kinetic control of cellular programs such as apoptosis. To test this hypothesis, one needs a systematic titration of the effects of intrinsic helicity on the mechanisms of coupled folding and binding. There is no easy way to modulate intrinsic helicities for an IDP that adopts helical conformations in its bound state. Mutagenesis experiments inevitably convolve changes to amino acid composition and intrinsic helicities, as is the case with standard, proline-, glycine- or alanine-scanning approaches. This makes it difficult it to separate the contributions of intrinsic helicities from the overall effects of changes to the amino acid composition. In this regard, it is noteworthy that the amino acid compositions and residues that define macromolecular interfaces are highly conserved in IDPs even though their amino acid sequences vary considerably [21, 22]. Our goal is to develop an approach that allows us to parse contributions from amino acid composition and sequence-encoded intrinsic

helicities in order to uncover their distinct and synergistic contributions to thermodynamic and kinetic stabilities of complexes that form via coupled folding and binding. Here, we present a method that we refer to as GADIS for **G**enetic **A**lgorithm for the **D**esign of **I**ntrinsic secondary **S**tructures. This approach combines a genetic algorithm and efficient molecular simulations to design IDP sequences that have specified helicity profiles in their unbound forms.

In the implementation of the GADIS algorithm that is presented here, we take a position-specific helicity profile and two additional sets of constraints as inputs. The constraints are as follows: We fix the amino acid composition thus eliminating the need for traditional proline or alanine scanning methods that change the amino acid composition. We also fix the positions of residues that define the interface of the IDP with its binding partner. The goal is to design a set of sequences that reproduces the target helicity profile for the given amino acid composition. We have prototyped GADIS by using it to generate sequence variants of the 34-residue IDR within PUMA that binds to Mcl-1. We show that GADIS is successful and efficient at generating distinct sequence variants that satisfy specific design criteria for helicity profiles. We report results from far ultraviolet circular dichroism (UV-CD) measurements for ten of the designed sequence variants, with different target helicity profiles and mean helicities. Quantitative comparisons show that computationally derived mean helicities are in agreement with those derived from experiment.

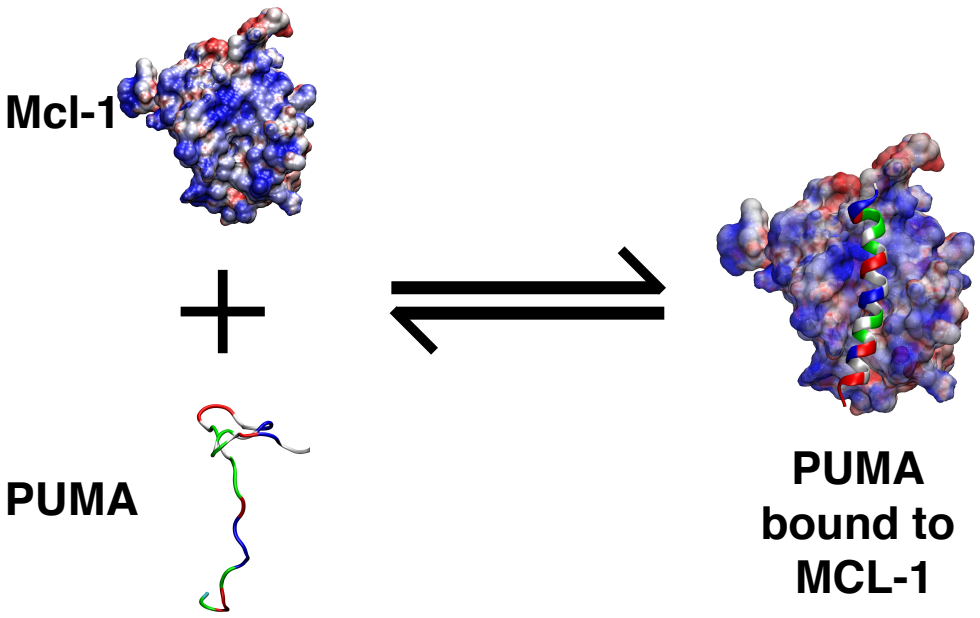


Figure 2.1: Illustration of coupled folding and binding. In this illustration, an intrinsically disordered – partially helical – PUMA sequence is shown to bind to Mcl-1 and form a continuous helix in the context of the bound complex. PUMA is shown as a ribbon diagram to emphasize its helicity in the bound complex. The residues are colored as follows: Hydrophobic residues are in gray, polar residues are in green, negatively charged residues are in red, and positively charged residues are in blue. Mcl-1 is shown in a surface representation to emphasize the electrostatic potential. Regions of high positive potential are in blue, regions of high negative potential are in red, and regions with near zero electrostatic potential are in white. The electrostatic surface was computed using the Adaptive Poisson Boltzmann solver [23].

2.2 GADIS Algorithm

We illustrate the design objectives and the functionality of GADIS using PUMA. The wild type version of PUMA adopts a continuous alpha helix in the context of its complex with Mcl-1 (Figure 2.1). In its unbound state, PUMA adopts a heterogeneous ensemble of partially helical conformations (Figure 2.2). This translates to a residue-specific helicity profile (Figure 2.2) that quantifies the ensemble-averaged percent probability of finding each residue as part of a regular alpha helical segment of at least six consecutive residues.

The flowchart in Figure 2.3 illustrates the steps involved in GADIS. The algorithm involves two initialization steps I1 and I2. In **step I1** we specify the inputs, which include the amino acid

composition, the positions and identities of immutable residues, and the target helicity profile. In **step I2**, we start with the wild type sequence and generate 100 distinct seed sequences. For the first iteration, the algorithm segues directly into **step 3** of the production run. Here, for each seed sequence, we perform preliminary atomistic Metropolis Monte Carlo simulations based on the ABSINTH implicit solvation model and forcefield paradigm (see Methods section). Each simulation involves 3×10^7 steps that follow 10^7 initial steps of equilibration. The simulations yield conformational ensembles for each seed sequence. In **step 4**, the simulated ensembles are used to

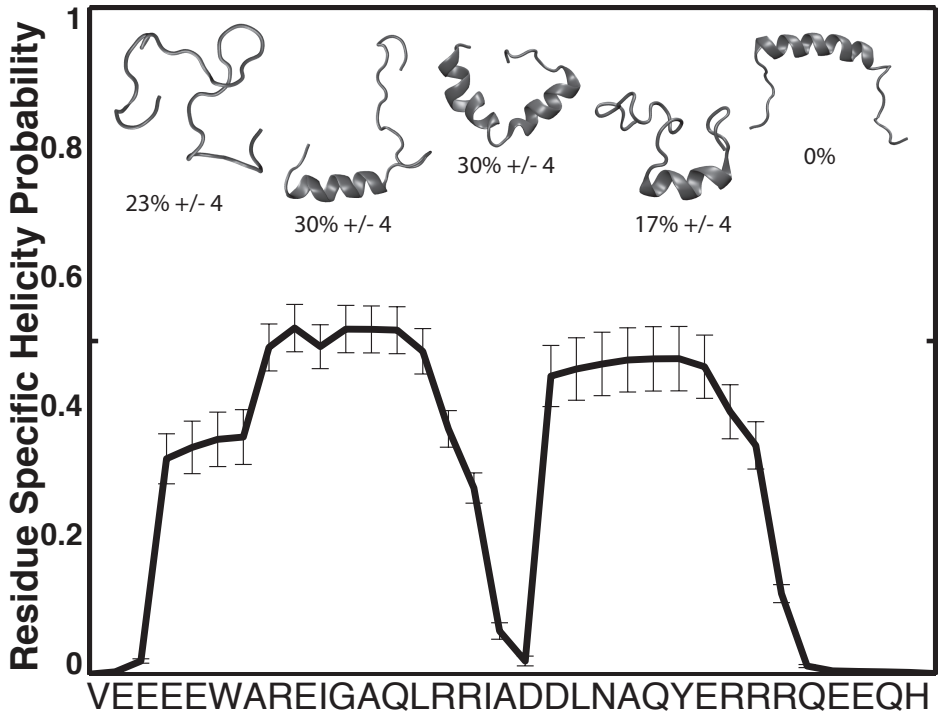


Figure 2.2: The unbound PUMA adopts a heterogeneous conformational ensemble. The figure summarizes results from all atom ABSINTH-based simulations of PUMA. The sequence prefers a heterogeneous ensemble of conformations. These include conformations with independent N- and C-terminal helical halves, coil-like N- or C-terminal halves that are populated with helical C- or N-terminal halves, and fully coil-like conformations. The heterogeneity is quantified in terms of the percent probabilities associated with distinct conformational types. These populations are used to quantify a residue-specific helicity profile that quantifies the percent probability of finding a residue as part of a regular alpha helical segment that is at least six residues long. Note that in the simulations the central helix conformation is not accessed by the wild type sequence of PUMA.

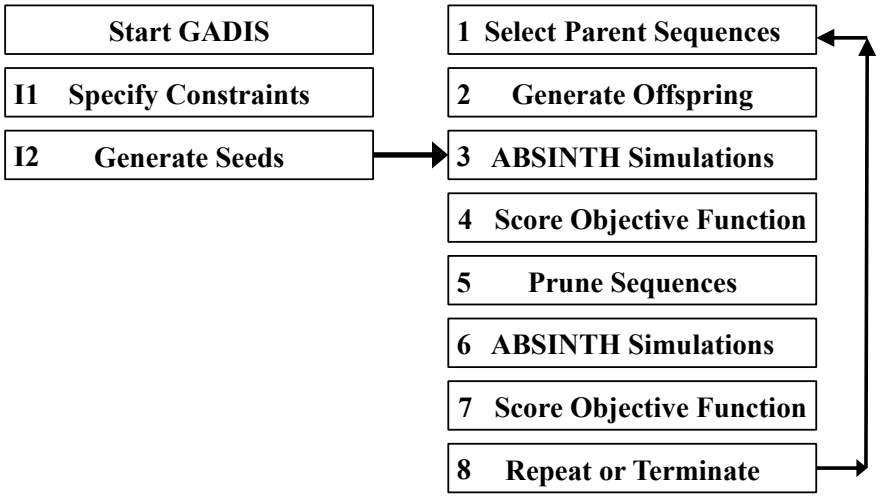


Figure 2.3: Flowchart of the GADIS algorithm. The text provides a detailed description of each of the steps in the algorithm.

calculate sequence-specific values of the objective function shown in equation (1). This quantifies the distance between the profile achieved by the conformational ensemble of each sequence and the target helicity profile. The objective function is defined as follows:

$$\Omega_k = \frac{1}{N}\sum_{i=1}^{N} w_i \left(p_{h,i}^{s,k} - p_{h,i}^{t,k} \right)^2 ; \qquad (1)$$

Here, Ω_k is the objective function for the k^{th} sequence, N is the number of residues in each sequence, $p_{h,i}^{s,k}$ is the percent probability of finding residue i in a helical segment of at least six residues within the simulated ensemble, and $p_{h,i}^{t,k}$ is the target value for this percent probability. The parameters w_i define the contribution of each position to the target helicity profile. These can either be uniform or non-uniform. The latter choice is useful if a specific target helicity profile has degeneracy. This refers to a similar Ω_k value being achieved by a range of distinct helicity profiles, including those that deviate from the intended target. The choices for w_i are made following initial testing, which allows us to assess the ease of generating sequences that match the target helicity profile. The assessments in **step 4** are used in **step 5** to prune the number of seed / parent sequences. This

pruning is achieved by selecting ten of the 100 original sequences with the lowest values of Ω_k. For the subset of selected sequences, we perform, in **step 6**, an additional round of ABSINTH-based Monte Carlo simulations, whereby ten independent simulations, each of length 4×10^7 steps are performed for each sequence. These simulations provide robust statistics that are used for evaluating the probability that a seed sequence can be used as a parent for generating offspring sequences in the next generation. Specifically, the conformational statistics are used to calculate a new round of objective function values, and the seed sequences are evaluated for their potential to become parents for the next generation of sequences in **step 7**. If at least ten distinct sequences have been generated that match the target helicity profile and the best set of sequences have not improved over the last two generations, then the design process is terminated. If these criteria have not been met, then new offspring sequences are to be generated and the design continues whereby we return to **step 1** and iterate **steps 1 – 7** until the termination criterion has been satisfied. In our tests with PUMA, the GADIS procedure typically yields the desired number of sequence variants within eight generations and this is true irrespective of the target helicity profile.

The details of selecting parent sequences, **step 1**, and generating offspring sequences, **step 2**, are as follows: In **step 1**, the probability P_k that an offspring sequence will be derived from parent sequence k is given by:

$$P_k = \frac{\exp\left(-c\Omega_k\right)}{\sum_{k'=1}^{n_p} \exp\left(-c\Omega_{k'}\right)}; \tag{2}$$

Here, n_p represents the current number of parent sequences including any that seeded the previous generations. The choice for c that is currently used for designing variants of PUMA is shown in equation (3):

$$c = \frac{12N}{\sum_{i=1}^{N} w_i} \; ; \tag{3}$$

This value of *c* works well in terms of affording an efficient balance between sequence diversity and achievement of the target profile in the choice of parent sequences. The new set of parent sequences and parents from the preceding generations are used to generate 100 new offspring sequences in **step 2**. From a parent sequence, offspring sequences are generated by swaps between pairs of residues at mutable positions as shown in Figure 2.4. Additional sliding moves alter the current positions of residues as shown in Figure 2.4. The swaps and slides are guided by positive and negative selection heuristics. The negative selection heuristics refer to biases against the accumulation of acidic / basic residues at C-terminal / N-terminal ends of helical segments. Additional criteria refer to biases against the inclusion of glycine or proline residues within internal helical segments of a sequence unless this is required by the input constraints. The positive selection heuristics are based on rules regarding helix initiation and capping. Residues that are known to be preferred at N- or C-termini of helices are preferentially chosen to be at these positions providing these choices are permitted by the fixed amino acid composition [24].

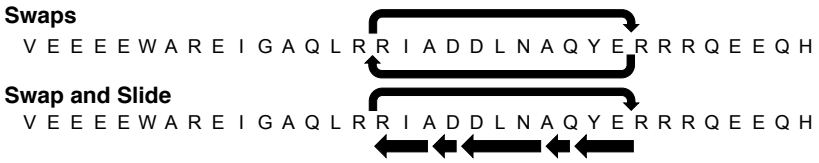


Figure 2.4: Illustration of the shuffles and sliding moves along sequences that are used to generate new offspring sequences from a parent. The top row illustrates swaps between two positions and the bottom row illustrates a combination of swaps and sliding. The latter to refers to changes to the positions of residues by sliding them over either to N- or C-terminal positions. Note that in the swap and slide move that the longer arrows signify a residue being moved over an immutable residue.

2.3 Deployment and Analysis of the Performance of GADIS

We prototyped GADIS by generating sequence variants of PUMA. The helicity profile for the wild type sequence is shown in Figure 2.2. We proposed five distinct target profiles for new variants of PUMA. These targets are shown in Figure 2.5. In **Target 1** the goal was to design sequences whose N- and C-terminal halves fluctuate independently into and out of helical conformations, with a clear break in the middle of the sequence. This target was referred to as the stable broken helix (SBH) profile. In **Target 2** the goal was to design sequences where a stable central helix spans the central portion of the peptide from positions 10-23. This target was referred to as the stable central helix (SCH) profile. In **Targets 3 and 4**, the goal was to design sequences that have helical N- or C-terminal halves and coil-like C- or N-terminal halves, respectively. These targets were referred to as NTH and CTH profiles, respectively. Finally, for **Target 5**, the goal was to achieve sequences with uniformly low probabilities of being part of regular helical segments. This target was referred to as the uniformly unstable helix (UUH).

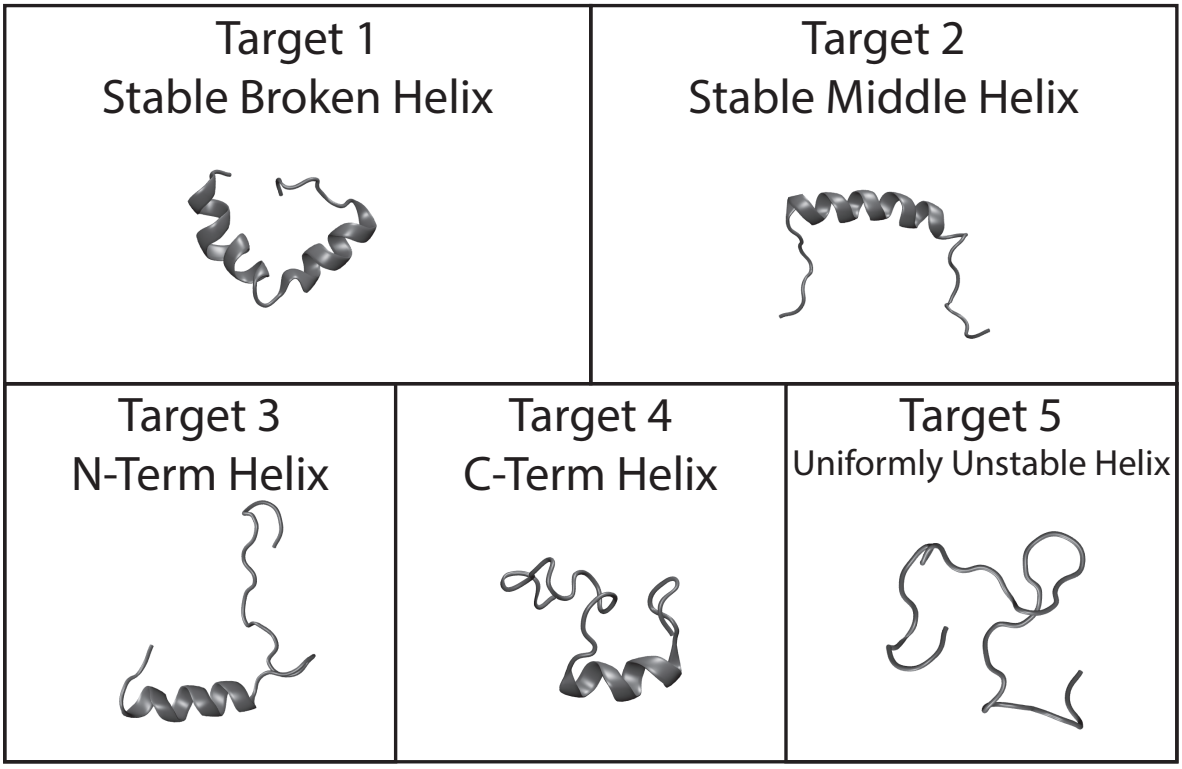


Figure 2.5: Five target helicity profiles for the design of PUMA variants. The acronyms and the details regarding each target profile are discussed in the text.

Figures 2.6 and 2.7 summarize the results of applying GADIS to generate at least ten distinct sequence variants for each of the five target helicity profiles. In these figures, the results are summarized as checkerboard plots that quantify the percent probabilities that each residue in a designed sequence is part of a regular alpha helical segment that is at least six residues long. The sequences that match a specific target profile are also shown adjacent to the checkerboard plots. Targets such as the SCH profile will be more challenging because this profile calls for persistent helicity across the central portion of the sequence with coil-like dangling ends. From a computational standpoint, the constraints of fixed amino acid composition and seven immutable positions present one set of challenges for the efficient generation of parent / offspring sequences that match the target helicity profile. An additional challenge comes from the degeneracy of incorrect helicity profiles that reproduce low Ω_k values for the SCH profile. This latter challenge is

remedied by using non-uniform weights w_i to prevent sequences encoding the SBH profile from generating low Ω_k values when the SCH profile is the intended target. In contrast, the UUH target is easily achieved by almost any sequence that is chosen at random. Figure 2.8 shows how the GADIS algorithm improves from one generation to the next by increasing the probability of finding sequence variants of PUMA that lower the value of Ω_k for the SBH profile. Similar results are obtained for each of the other four profiles.

Figure 2.6: Sequence variants of PUMA that were generated using GADIS for the SBH and SCH target profiles. The checkerboard plots quantify the residue-specific helical propensities. These are quantified in terms of the percent probability that a specific residue is part of a regular helical segment that is at least six residues long. On the left, the first ten rows pertain to sequence variants that correspond to the SBH profile and the bottom ten rows correspond to the SCH profile. The sequences corresponding to each row of residue-specific helical propensities are shown on the right. These positions of the immutable residues are highlighted to emphasize the constraints. The wild type PUMA sequence is also shown as reference. Additionally, sequences shown in bold face were used in UV-CD measurements.

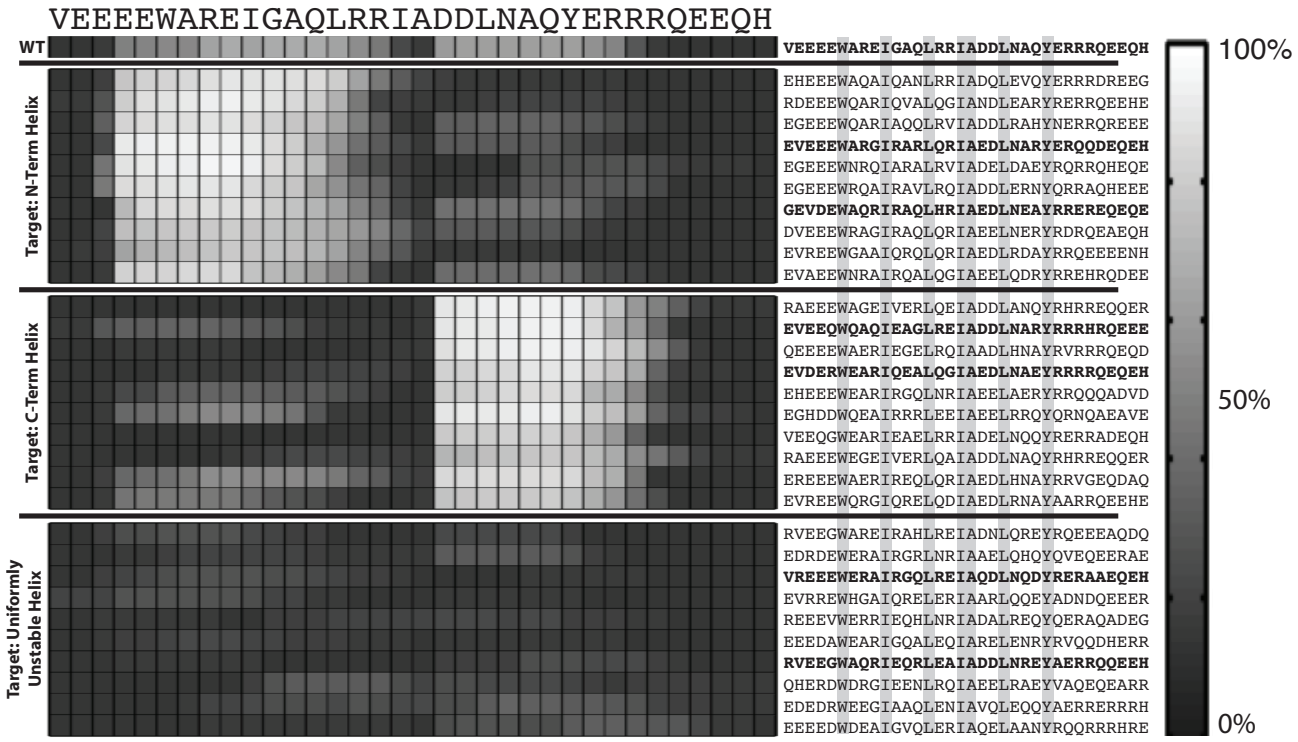


Figure 2.7: Sequence variants of PUMA that were generated using GADIS for the NTH, CTH, and UUH target profiles. The checkerboard plots quantify the residue-specific helical propensities. These are quantified in terms of the percent probability that a specific residue is part of a regular helical segment that is at least six residues long. On the left, the first ten rows pertain to sequence variants that correspond to the NTH profile, the middle ten rows correspond to the CTH profile, and the last ten rows correspond to the UUH profile. The sequences corresponding to each row of residue-specific helical propensities are shown on the right. These positions of the immutable residues are highlighted to emphasize the constraints. Additionally, sequences shown in bold face were used in UV-CD measurements. The wild type PUMA sequence is also shown as reference.

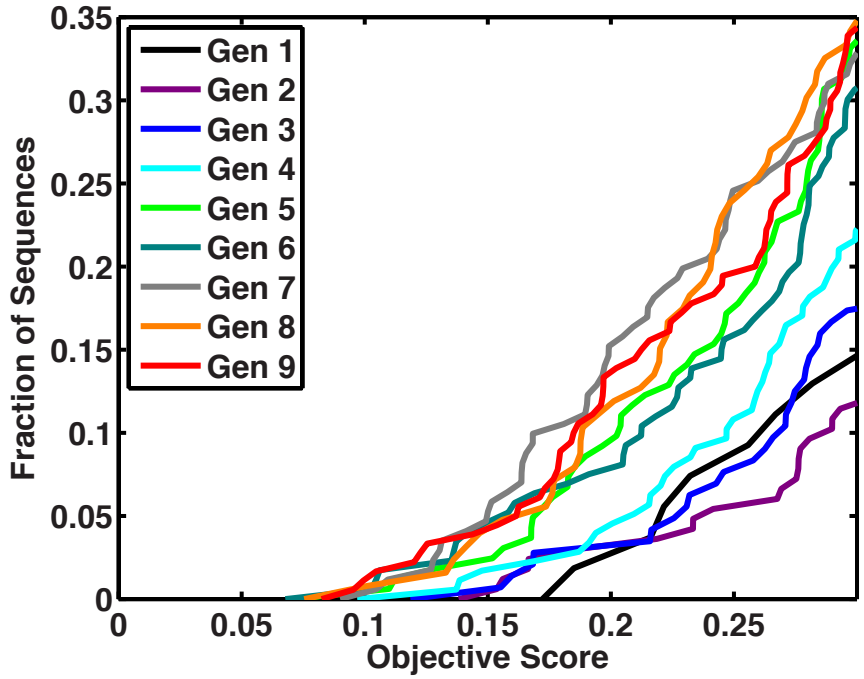


Figure 2.8: Quantifying the convergence of the GADIS algorithm. This plot shows the probability of realizing sequences with lower objective function values as the generation number increases. For a given curve, the ordinate quantifies the fraction of sequences generated by GADIS that have achieved a sequence with a score that is less than or equal to a particular value along the abscissa. As the generation number increases (see legend), the curves are shifted to the left indicating a systematic improvement in realizing sequences that lower the objective function value.

2.4 Experimental Validation of GADIS Results

We performed UV-CD measurements on ten different sequence variants, two from each of the five target classes. We also measured the CD spectrum of wild type PUMA. Figure 2.9 shows the CD spectra for all eleven sequences. We compared the calculated mean helical contents for wild type PUMA and each of the ten designed variants to the measured helical contents. For sequence k the mean helical content $f_{h,k}^{\text{calc}}$ is calculated using the residue-specific probabilities that are extracted from the simulated ensembles:

$$f_{h,k}^{\text{calc}} = \frac{1}{N}\sum_{i=1}^{N} p_{h,i}^{s,k} \quad ; \tag{4}$$

The values obtained using equation (4) were compared to mean helical contents inferred from analysis of the measured CD spectra, which was calculated using the empirical equation developed by Chen et al. [25]:

$$f_{h,k}^{\exp} = \frac{\theta_{222}}{3.95\times10^{4}\left(1-\frac{2.57}{N}\right)} \tag{5}$$

Here, θ_{222} is the mean residue ellipticity at 222 nm and N=34 is the number of amino acids in the sequence. The denominator is the expected mean residue ellipticity at 222 nm, calculated for an infinitely long helix and corrected to account for the finite size of the peptide. Other empirical expressions have also been developed that use either θ_{222} [26] or θ_{208} [27], which is the mean residue ellipticity at 208 nm. These expressions yield similar estimates for the inferred values, and identical trends, for mean helicities given our CD data.

Figure 2.10 shows a comparison between the values of $f_{h,k}^{\text{calc}}$ and $f_{h,k}^{\exp}$ for wild type PUMA and all ten designed variants derived from the application of GADIS. The two sets of values are positively correlated, although $f_{h,k}^{\text{calc}} \neq f_{h,k}^{\exp}$. This could derive from the discrepant approaches for estimating helicities, the parameterization of $f_{h,k}^{\exp}$ in equation (5), or true deviations in the ensembles sampled computationally versus in solution. Overall, we conclude that the GADIS designs do indeed enable a systematic titration of helicity profiles and mean helicities while maintaining the overall amino acid composition and fixing the positions of several immutable residues.

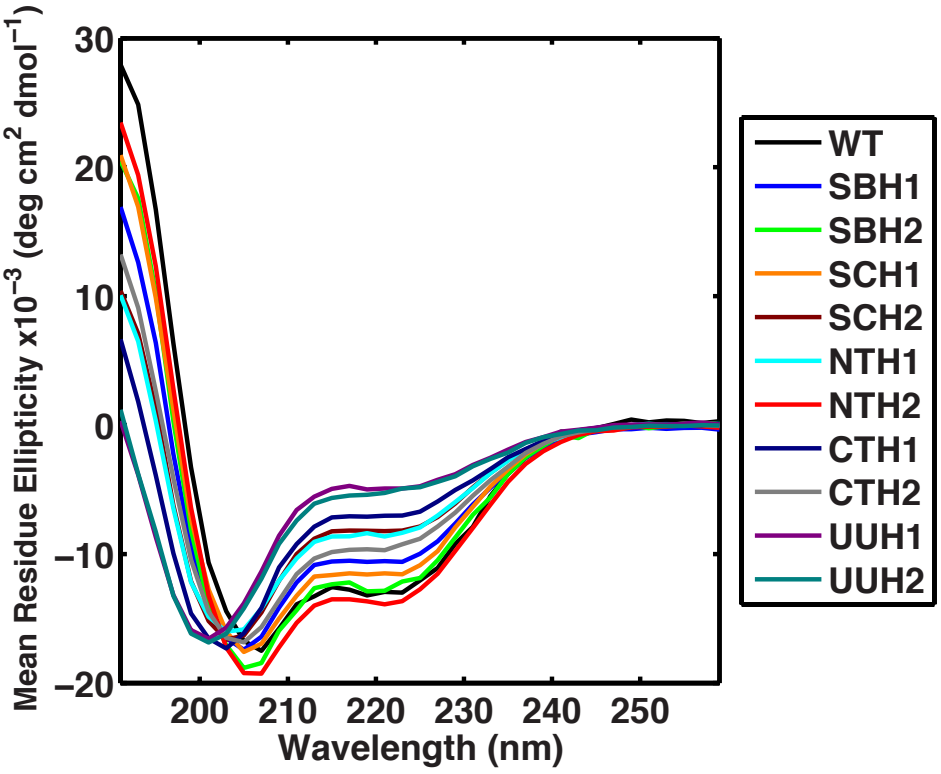


Figure 2.9: UV-CD spectra obtained for the wild type PUMA and ten sequence variants derived from the GADIS designs. The spectra show that GADIS helps achieve a systematic titration of intrinsic helicities through sequence design using a fixed amino acid composition and a specified set of immutable residues.

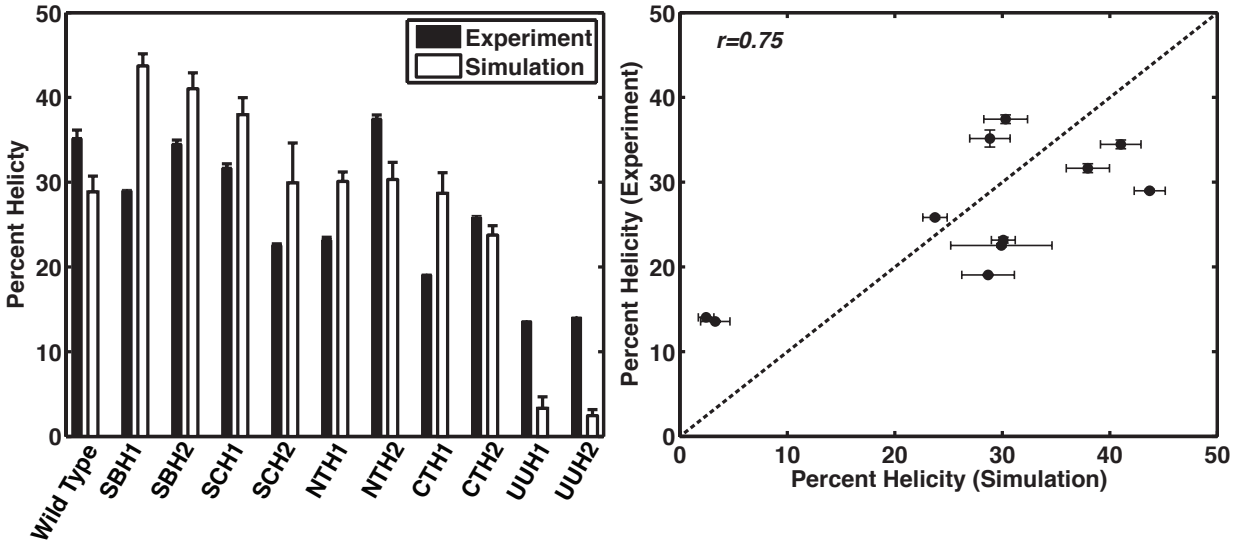


Figure 2.10: Comparisons between measured and calculated mean helical contents. The plot on the left shows the comparisons as a bar plot, where the black bars denote mean helical contents derived from CD spectra and the white bars denote the corresponding values derived from simulated ensembles for each sequence. The panel on the right plots the experimentally derived values on the ordinate versus the computationally derived values on the abscissa. The Pearson product moment correlation coefficient is $r = 0.75$ and this quantifies the linear correlation between the mean helical contents derived from measurements versus simulations. The p-value is 0.007 and this quantifies the probability of realizing the obtained r-value purely by chance. In the plot on the right, if the computed helicities were identical to the measured helicities, then the points would have fallen on the dashed line. The vertical error bars are the differences between the helicity values inferred from the two sets of experiments. The horizontal error bars represent the standard error about the mean helicity that is calculated across at least ten independent simulations for each sequence variant.

2.5 Why use ABSINTH-based simulations?

In **step 3** and **step 6** of the GADIS algorithm we use ABSINTH-based simulations to generate atomistic descriptions of conformational ensembles to calculate sequence-specific helicity profiles. This is the most computationally expensive step of the GADIS algorithm. For a typical sequence variant of PUMA, it takes roughly 48 hours to complete a simulation on a quad core Nehalem processor. This can become a major bottleneck given the need to return to steps 3 and 6 multiple times for hundreds of sequences. We overcome this problem through our access to a high performance computational cluster. This still requires at least 720 hours of continuous computations, and can become prohibitive without access to requisite resources.

The computational bottleneck raises the issue of finding inexpensive ways to estimate of sequence-encoded helicities. We used the ABSINTH-based approach based on previous work that uncovered limitations of web-based predictors of helicity such as AGADIR [28]. Although AGADIR is routinely used to estimate helicities of various peptides and proteins, it does not appear to capture the sequence-encoded intrinsic helicities of IDPs / IDRs [14]. This point is reinforced in Figure 2.11, which shows the poor correlation between helicities predicted using AGADIR and the values from simulations or the values of from UV-CD measurements for PUMA and the ten different sequence variants. Therefore, pending the availability of a suitable machine learning approach that can be deployed across a large dataset of sequences, we are constrained to using ABSINTH-based simulations at steps 3 and 6 of the GADIS algorithm. The efficiency of ABSINTH-based simulations enables the throughput in terms of the number of simulations and the realization of the design objectives. This would not have been feasible with the use of explicit representations of solvent molecules or an inefficient implicit solvation models.

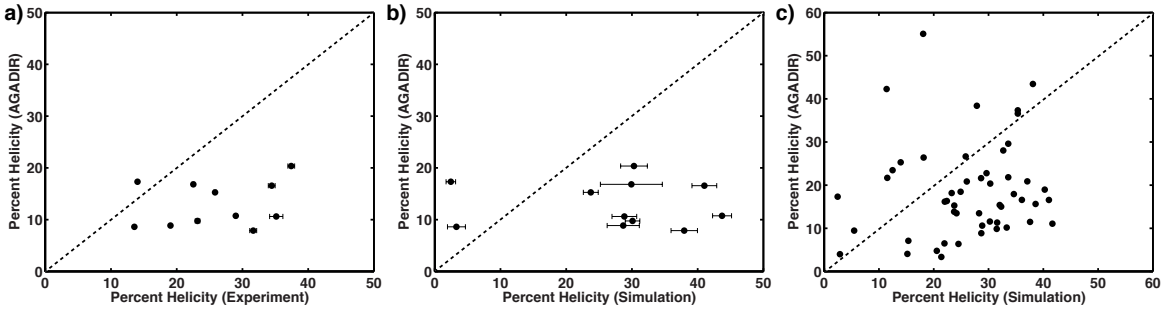


Figure 2.11: Comparisons between mean helical contents obtained using AGADIR and those derived from CD measurements (a) and simulations (b), (c). In all three panels, if the AGADIR values were identical to the values along the abscissae, then the points would fall on the dashed lines shown in each of the three panels. AGADIR predictions were performed using default settings for the ionic strength and a temperature of 25˚C. This yields uniformly low helicity values for all eleven sequences. It also fails to capture the variation of intrinsic helicities with sequence. Similar trends, albeit lower helicity values are obtained by setting a salt concentration of 108 mM and temperature of 298.15 K. For the plot in panel (a), $r = -0.07$ and $p = 0.85$ and for the plot in panel (b), $r = 0.23$ and $p = 0.49$. In panel (a), the horizontal error bars are the differences between the helicity values inferred from the two sets of experiments. In panel (b), the horizontal error bars represent the standard error about the mean helicity that is calculated across at least ten independent simulations for each sequence variant. Panel (c) shows a comparison of mean helicities derived from AGADIR versus those derived from the simulated ensembles for all fifty-one sequences shown in Figures 6 and 7. With five times more data than in panels (a) and (c), the data in panel (c) establish a consistent lack of correlation ($r = 0.1$ and $p = 0.48$) between AGADIR and ABSINTH-based mean helicities. These results are consistent with previous observations made on a different set of IDPs that show favorable comparisons between simulation results and experimental data and poor correlations when using AGADIR-based predictions [14].

2.6 Discussion

We have succeeded in developing and deploying a systematic titration of intrinsic helicity profiles while satisfying the two constraints that we imposed on our design strategy. Deploying these designs in mechanistic experiments should enable detailed investigations of the impact of changes to intrinsic helicity, given a fixed amino acid composition, on the mechanisms of coupled folding and binding of IDPs that adopt helical conformations in their bound complexes. Experiments to investigate the effects of GADIS-based designs of PUMA on the binding to Mcl-1 are currently underway. Insights from these experiments should pave the way for an iterative procedure of assessing the effects of fewer or larger number of constraints on the designs. These designs that achieve target helicity profiles, when coupled to binding data, will help us uncover the sequence and structural determinants of specificity in coupled folding and binding.

Currently, GADIS can be deployed to any design problem that fits the PUMA archetype, and there are several such problems in the coupled folding and binding field. Interestingly, there are also several problems in spontaneous unimolecular folding that are similar in spirit to the coupled folding and binding problem. The folding of linear repeat proteins is one such example [29]. Here, free energy of folding is governed by the interplay between the intrinsic instability of a repeat versus the favorable interfacial free energy between repeats [30]. GADIS, in its current form, can be deployed to redesign helical units in repeat protein to preserve the interfacial residues and amino acid compositions. This would enable a modulation of the balance between the intrinsic versus interfacial free energies and allow one to assess the impact of redesigns on overall stability and the cooperativity of folding. GADIS can also be generalized to work with fewer constraints on amino acid compositions or tightening the constraints in terms of specifying additional immutable residues that might contribute indirectly to stabilizing the interfaces between complexes. These generalizations of GADIS should be tailored to specific set of experiments that one has in mind since the algorithm has been developed to guide systematic sequence titrations that test specific hypotheses about intrinsic and coupling free energies.

2.7 Methods

All atom simulations: The simulations were performed using version 2.0 of the CAMPARI molecular modeling suite (http://camapri.sourceforge.net). This package provides full support for the ABSINTH implicit solvation model and forcefield paradigm [31]. In ABSINTH, the polypeptide chain and solution ions are modeled in atomistic detail. The solvent is modeled as a continuum that responds to conformational fluctuations through changes to atom-specific solvation states that modulate the reference free energies of solvation and solvent-mediated electrostatic interactions. All parameters for the forcefield were from the abs_3.2_opls.prm parameter file. Each

simulation was initialized using a randomly generated self-avoiding conformation and distinct random seed. We set the simulation temperature to be 310 K and performed Metropolis Monte Carlo simulations using standard move sets that were previously deployed for simulations of other IDRs with intrinsic helicities [14].

Design constraints and GADIS software: For PUMA, we use a numbering scheme that goes from 1 – 34. The overall amino acid composition is held fixed in the GADIS designs. All sequences were N-methylamidated at the N-terminus and acetylated at the C-terminus. Seven hydrophobic residues *viz*., W6, I10, L14, I17, A18, L21, and Y25 define the interfacial contacts between the folded PUMA sequence and Mcl-1. Accordingly, these seven are set as being immutable in the GADIS designs. This implies that their positions are held fixed and the identities are not changed when the swap / slide moves are deployed to generated offspring sequences. The implementation of heuristics that guide the GADIS-based design of offspring sequences is shown in the form of pseudo-code and is included as Figure 2.12. The evaluation of objective functions, the selection of parent sequences, and the generation of offspring sequences were implemented in MATLAB. The code was designed to interface with outputs from CAMPARI-based simulations.

UV-CD experiments: For the experiments, we purchased peptides with capped termini in pure form from Watsonbio Sciences. Mass spectrometry analysis from the vendor combined with amino acid analysis confirmed the identities of the peptides. All the peptides were reconstituted using 50 mM Sodium Phosphate pH 7.0, 0.05% (v/v) Tween 20. To remove residual salts, peptides were exchanged into 50 mM Sodium Phosphate pH 7.0, 0.05% (v/v) Tween 20 using HiTrap Desalting columns (GE Healthcare). The peptide concentrations for CD experiments were estimated using the absorbance measurements and use of Beer-Lambert law with an extinction coefficient of 7113 M^{-1} cm^{-1} at 280 nm. Final peptide stock concentrations were determined from the mean of two amino

acid analysis runs. The final concentrations for UV-CD measurements were small and in the range of 2.5-10 μM. Care was taken to ensure that the results of our measurements are not confounded by peptide oligomerization.

For the CD measurements, each peptide was prepared and scanned in a single day. Peptides were diluted individually from the stock by weight. Two samples were prepared for each concentration. At least three different concentrations were scanned and compared to check for concentration dependence. The two samples from the highest concentration of peptide that did not show concentration dependence were averaged to give the final mean residue ellipticity. CD scans were performed at 25 °C using an Applied Photophysics Chirascan and a 2 mm path length cuvette. Settings were 1 nm bandwidth and 15 s adaptive averaging. To rule out changes in signal as a function of time, separate measurements were performed following one-hour time intervals between the scans for each sample at the same concentration. The measured CD signal was converted to Mean Residue Ellipticity (MRE) by dividing through by the concentration (M), the cuvette path length (cm) and the total number of amino acid residues. For comparisons to computational results, the peptide MRE was reported as the mean of the highest concentration samples that did not display concentration dependence (3.5 μM for wild type, 5 μM for SBH2, and 10 μM for the remaining peptides).

2.8 Acknowledgments

The US-National Science Foundation and US-National Institutes of Health supported this work through grants MCB-1121867 and 5RO1 NS056114, respectively to RVP. JC and SLS were supported by the Wellcome Trust (WT 095195MA). MDC was supported by a Biotechnology and

Biological Sciences Research Council (BBSRC) studentship. The MATLAB code for aiding GADIS designs is available from the authors upon request.

2.9 References

1. Harmon, T.S., et al., *GADIS: Algorithm for designing sequences to achieve target secondary structure profiles of intrinsically disordered proteins.* Protein Eng Des Sel, 2016. **29**(9): p. 339-46.
2. van der Lee, R., et al., *Classification of Intrinsically Disordered Regions and Proteins.* Chemical Reviews, 2014. **114**(13): p. 6589-6631.
3. Wright, P.E. and H.J. Dyson, *Linking folding and binding.* Curr Opin Struct Biol, 2009. **19**(1): p. 31-8.
4. Wright, P.E. and H.J. Dyson, *Intrinsically unstructured proteins: Re-assessing the protein structure-function paradigm.* Journal of Molecular Biology, 1999. **293**(2): p. 321-331.
5. Babu, M.M., R.W. Kriwacki, and R.V. Pappu, *Versatility from Protein Disorder.* Science, 2012. **337**(6101): p. 1460-1461.
6. Wright, P.E. and H.J. Dyson, *Intrinsically disordered proteins in cellular signalling and regulation.* Nat Rev Mol Cell Biol, 2015. **16**(1): p. 18-29.
7. Dunker, A.K., et al., *Intrinsic disorder and protein function.* Biochemistry, 2002. **41**(21): p. 6573-6582.
8. Mohan, A., et al., *Analysis of molecular recognition features (MoRFs).* J Mol Biol, 2006. **362**(5): p. 1043-59.
9. Dyson, H.J. and P.E. Wright, *Coupling of folding and binding for unstructured proteins.* Current Opinion in Structural Biology, 2002. **12**(1): p. 54-60.

10. Gianni, S., J. Dogan, and P. Jemth, *Coupled binding and folding of intrinsically disordered proteins: what can we learn from kinetics?* Curr Opin Struct Biol, 2016. **36**: p. 18-24.

11. Peng, Z.L., et al., *A creature with a hundred waggly tails: intrinsically disordered proteins in the ribosome.* Cellular and Molecular Life Sciences, 2014. **71**(8): p. 1477-1504.

12. Vacic, V., et al., *Characterization of molecular recognition features, MoRFs, and their binding partners.* J Proteome Res, 2007. **6**(6): p. 2351-66.

13. Dyson, H.J. and P.E. Wright, *Intrinsically unstructured proteins and their functions.* Nature Reviews Molecular Cell Biology, 2005. **6**(3): p. 197-208.

14. Das, R.K., S.L. Crick, and R.V. Pappu, *N-Terminal Segments Modulate the alpha-Helical Propensities of the Intrinsically Disordered Basic Regions of bZIP Proteins.* Journal of Molecular Biology, 2012. **416**(2): p. 287-299.

15. Borcherds, W., et al., *Disorder and residual helicity alter p53-Mdm2 binding affinity and signaling in cells.* Nat Chem Biol, 2014. **10**(12): p. 1000-1002.

16. Kiefhaber, T., A. Bachmann, and K.S. Jensen, *Dynamics and mechanisms of coupled protein folding and binding reactions.* Curr Opin Struct Biol, 2012. **22**(1): p. 21-9.

17. Rogers, J.M., A. Steward, and J. Clarke, *Folding and binding of an intrinsically disordered protein: fast, but not 'diffusion-limited'.* J Am Chem Soc, 2013. **135**(4): p. 1415-22.

18. Rogers, J.M., et al., *Interplay between partner and ligand facilitates the folding and binding of an intrinsically disordered protein.* Proc Natl Acad Sci U S A, 2014. **111**(43): p. 15420-5.

19. Rogers, J.M., C.T. Wong, and J. Clarke, *Coupled folding and binding of the disordered protein PUMA does not require particular residual structure.* J Am Chem Soc, 2014. **136**(14): p. 5197-200.

20. Dogan, J., et al., *Binding Rate Constants Reveal Distinct Features of Disordered Protein Domains.* Biochemistry, 2015. **54**(30): p. 4741-50.

21. Brown, C.J., et al., *Evolution and disorder.* Current Opinion in Structural Biology, 2011. **21**(3): p. 441-446.

22. Moesa, H.A., et al., *Chemical composition is maintained in poorly conserved intrinsically disordered regions and suggests a means for their classification.* Molecular BioSystems, 2012. **8**(12): p. 3262-3273.

23. Baker, N.A., et al., *Electrostatics of nanosystems: application to microtubules and the ribosome.* Proc Natl Acad Sci U S A, 2001. **98**(18): p. 10037-41.

24. Aurora, R. and G.D. Rose, *Helix capping.* Protein Sci, 1998. **7**(1): p. 21-38.

25. Chen, Y.H., J.T. Yang, and K.H. Chau, *Determination of the helix and beta form of proteins in aqueous solution by circular dichroism.* Biochemistry, 1974. **13**(16): p. 3350-9.

26. Chen, Y.H. and J.T. Yang, *A new approach to the calculation of secondary structures of globular proteins by optical rotatory dispersion and circular dichroism.* Biochem Biophys Res Commun, 1971. **44**(6): p. 1285-91.

27. Greenfield, N. and G.D. Fasman, *Computed circular dichroism spectra for the evaluation of protein conformation.* Biochemistry, 1969. **8**(10): p. 4108-16.

28. Lacroix, E., A.R. Viguera, and L. Serrano, *Elucidating the folding problem of alpha-helices: Local motifs, long-range electrostatics, ionic-strength dependence and prediction of NMR parameters.* Journal of Molecular Biology, 1998. **284**(1): p. 173-191.

29. Aksel, T. and D. Barrick, *Analysis of repeat-protein folding using nearest-neighbor statistical mechanical models.* Methods Enzymol, 2009. **455**: p. 95-125.

30. Aksel, T., A. Majumdar, and D. Barrick, *The contribution of entropy, enthalpy, and hydrophobic desolvation to cooperativity in repeat-protein folding.* Structure, 2011. **19**(3): p. 349-60.

31. Vitalis, A. and R.V. Pappu, *ABSINTH: A New Continuum Solvation Model for Simulations of Polypeptides in Aqueous Solutions.* Journal of Computational Chemistry, 2009. **30**(5): p. 673-699.

Chapter 3

Intrinsic Disorder That Isn't – Charge Patterning Contributes to the Formation of Stable Single Alpha Helices Through Preferential Charge Neutralization

This chapter is adapted from an article under preparation. Ammon E. Posey and Rohit V. Pappu designed and conducted the experiments. Tyler S. Harmon and Rohit V. Pappu developed the simulation framework. Tyler S. Harmon performed and analyzed the simulations.

3.1 Introduction

The intrinsically disordered proteome is enriched in polyampholytic sequences[1, 2]. Symmetric polyampholytes are enriched in charged residues and have equal or nearly equal numbers of positive and negative charges within the sequence. Of course, this charge counting rests on the assumption that there are no charge regulation effects and that the pK_a values of titratable groups remain fixed at their model compound values. If we make this assumption and perform simulations using the resultant "fixed charge" model, then IDPs would be simulated by setting Arg and Lys to have a net positive charge, Glu and Asp to have a net negative charge, and His to be neutral under conditions where the simulation conditions are intended to mimic a solution pH of ~ 7.0 – 7.4. The explicit inclusion of solution ions, modeled using the accurate and transferable parameters developed by Albert Mao, a previous graduate student in the lab, allows us to account for charge renormalization effects[3]. This refers to counterion condensation that neutralizes excess charge along polyions and ion-ion correlation effects that

even engender charge inversion when we think of the polymer and its condensed layer of ions. In strong polyampholytes, these effects are rather small because the sequences are already charge-balanced. There are two other charge-mediated effects namely, charge transfer, especially among pi systems, and charge regulation, which refers to the alteration of charge states of titratable groups via shifted pK_a values. To date, there has been an absence of systematic consideration of charge regulation effects in IDPs.

This work is motivated by a rather striking and seemingly paradoxical observation that forced us to revisit our assumptions about how we model IDPs. Das and Pappu invested considerable effort into modeling the effects of the patterning of oppositely charged residues on the overall dimensions, shapes, and amplitudes of conformational fluctuations of numerous archetypal IDPs[1]. The work was initiated by investigations of so-called EK-permutants, which was a set of sequences that pattern 25 distinct Glu and Lys residues differently in each of the 33 sequences studied. The finding was that segregating the Glu and Lys residues into blocks leads to significant chain compaction, whereas mixing the Glu and Lys residues along the linear sequence leads to dimensions that are considerably larger than even self-avoiding random walks. Das & Pappu developed a scaling theory to describe the observed behavior and recently Sawle and Ghosh developed a variational theory that also captures the observations regarding the impact of charge patterning on overall dimensions[4]. Within the basis set of sequences was a well-mixed sequence of the form $[(Glu)_4\text{-}(Lys)_4]_6$ which was predicted to form expanded coil-like ensembles. This prediction seemed to be consistent with the balancing of local charge segregation vs. global mixing of the 4-residue blocks. In these simulations that are based on the use of a fixed charge model, the preferential solvation of charged residues, combined with the screening of any local electrostatic attractions by other local and non-local electrostatic

repulsions engenders the observed coil-like behavior. In the question, do these predictions stand up to scrutiny, the answer is no.

Sequences based on repeats of the $(Glu)_4$-$(Lys)_4$ motif are actually quite prominent in nature[5-7]. They are especially common in myosin motors, where they are conjectured to serve as molecular 2×4's that connect globular domain and enable the requisite force generation during the power stroke of the motor[8]. This conjecture is based on the observation that repeats of $(Glu)_4$-$(Lys)_4$, referred to hereafter E4K4r sequences, actually form rigid alpha helices in solution[9]. This result stands in striking contrast to the results obtained based on fixed charge simulations that use the ABSINTH implicit solvation model and forcefield paradigm[1]. A readymade explanation for the experimental observations is that the model errs in capturing helix stabilizing intramolecular salt bridges. This is a perfectly reasonable conjecture, but it seems too simplistic for two reasons: In sequences such as E4K4r, there are numerous equivalent salt bridges to be made and in a purely additive model, there is a significant entropic penalty to overcome in terms of the degeneracy of conformations with equivalent numbers of intra-chain salt bridges. Secondly, there is a significant crowding of the solvation shells of oppositely charged residues that accompanies helix formation and it is not clear if salt bridges within a helix can overcome the desolvation penalty and the sharing of distinct types of solvation shells.

The results obtained using the ABSINTH model for $(E4K4)_3$ is shown in figure 3.1a. It is known that salt bridges in explicit solvent are over stabilized and the protein:solvent interaction is underestimated[10, 11]. These will bias the explicit solvent simulations toward favoring helical conformations with salt bridges at the expense of solvation. In examining helical conformations in the ABSINTH force field it becomes clear why the coil state is preferred over the helical state: the side chains do not have enough room in the helical conformation to maintain

their interactions with the surrounding solvent (Fig. 3.1b). Based on this observation and given the experimental data, we reasoned that the E4K4r system might be a perfect candidate for charge regulation whereby the changes to protonation states might engender a disorder-to-order transition. This conjecture was driven by the observation that the free energies of solvation of protonated Glu residues (designated as e) and deprotonated Lys residues (designated as k) are an order of magnitude smaller than their charged counterparts. This translates to a considerably smaller desolvation penalty and minimized overlap between the solvation shells of Glu and Lys residues along a helix as pictured in figure 3.1b. We reasoned that the source of the discrepancy between ABSINTH simulations and experiments in terms of the false positive assignment of disorder to E4K4r sequences must arise from charge regulation that is quenched in our simulations and poorly appreciated in the IDP field. This chapter tests this hypothesis using a repurposing of the GADIS algorithm and a testing of predictions that we obtain using potentiometric measurements[12]. We find evidence for preferential neutralization of Glu residues driving the formation of stable alpha helices. This stands in contrast to the salt bridge hypothesis. We show that simulations with neutralized Glu residues are consistent with the previous studies. We also make predictions regarding the identities of internal Glu residues that are most likely to be neutralized at neutral pH and show preliminary evidence, based on potentiometric titrations, which support our predictions regarding helix stabilization via preferential neutralization of internal Glu residues.

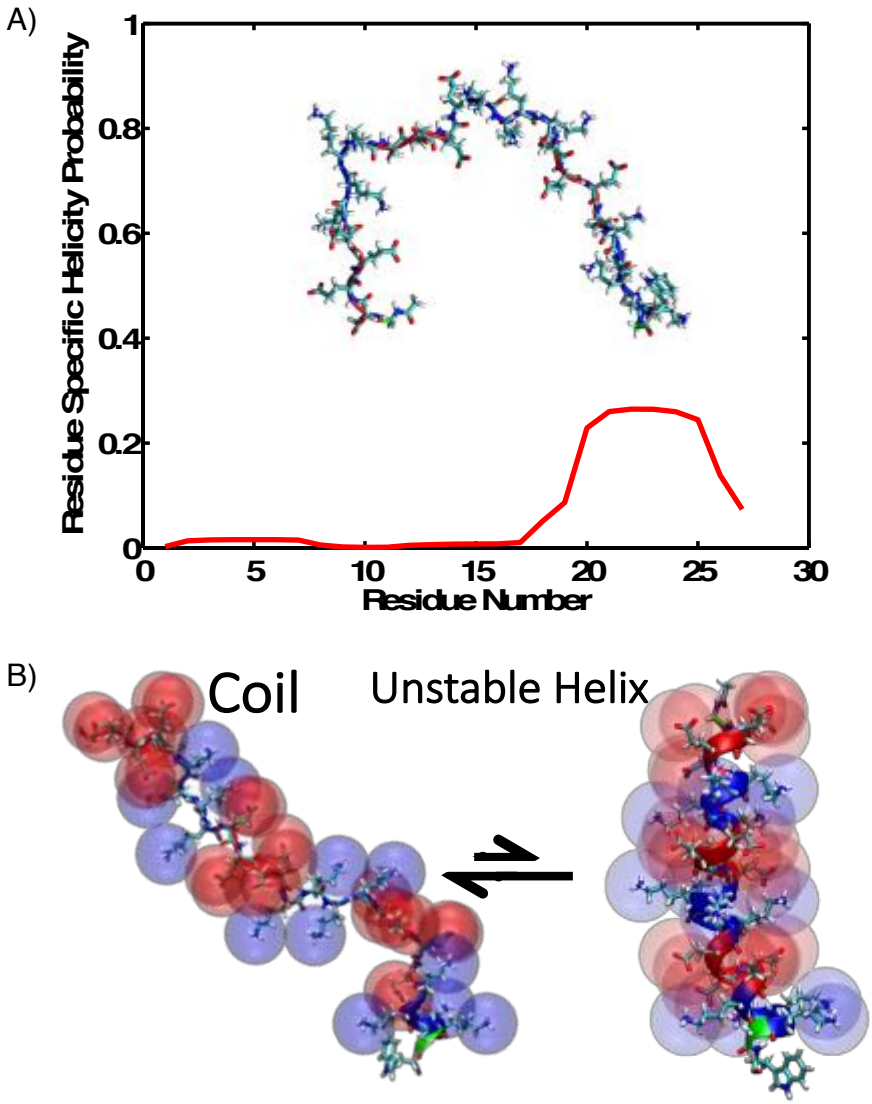


Figure 3.1: Absinth forcefield does not form stable helical conformations in the model compound charge state. (A) The residue resolved helicity from simulations with a representative snapshot of the protein in its random coil ensemble. (B) A cartoon representation to illustrate the difficulty of simulations forming a helical conformation. The protein conformations come from the unbiased simulation (coil) and a heavily biased simulation (helix). The colored spheres are shown to illustrate the solvation shells of the charged residues in the simulation and the difficulty in satisfying the solvation shells while in the helical conformation.

3.2 Deployment of GADIS

To explore the possibility of charge neutralization driving helix formation we deployed the GADIS algorithm to study the coupling between charge regulation and helix formation for the sequence $G(E4K4)_3GW$. This model peptide was recently studied in detail by the Woolfson lab[9]. Here, we used the ABSINTH forcefield, and the sequence shuffle moves in GADIS were replaced with moves that randomly protonate / deprotonate residues[12, 13]. The target function was a fully formed single alpha helix and the achievement of the target was monitored using an objective function that maximizes the distance between the current ensemble and the target structure. This function takes the form:

$$W = 1 - \frac{\sum_{i=1}^{N} h_i^2}{N}; \tag{3.1}$$

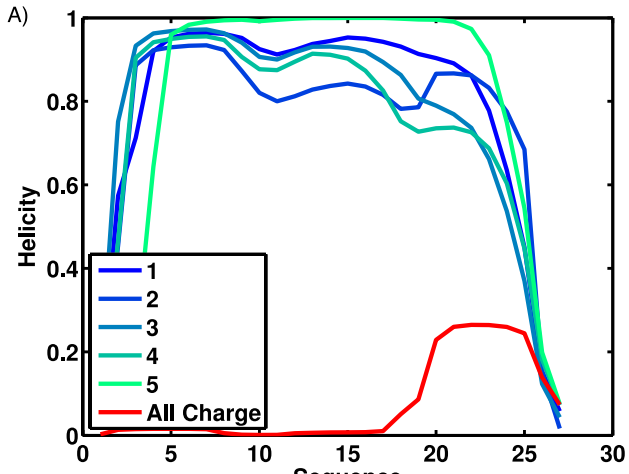
Here, h_i is the probability that residue I is part of a regular alpha helix and N is the number of residues in the sequence. GADIS was deployed to uncover sequences with neutralized Glu or Lys residues that spontaneously form stable alpha helices in contrast to residues with fully charged Glu and Lys residues.

3.3 Preferential Neutralization of Internal Glu Residues Leads to Stable Single Alpha Helices

Given the target of a stable single alpha helix, we found that GADIS converged very rapidly, within two generations, to yield a collection of sequences with preferential neutralization of Glu residues that lead to stable helical conformations using ABSINTH. We have plotted the residue resolved helicity for the top five scoring sequences, figure 3.2a. Of note is that, consistent with

the NMR data, these sequences with selective neutralization of Glu residues, states form a single alpha helix that spans the entire protein without breaking. In figure 3.2b we show the convergence of GADIS, which is quantified by the cumulative total number of sampled charge states that that have a given objective score or better for each generation. The third generation doing almost exactly as well as the second generation is taken as convergence of the algorithm.

In figure 3.2c we show the top five sequences with the neutralized residues highlighted as lowercase and black. Several key features become immediately obvious. First, out of the total 60 Lys residues in the top five sequences, only two are neutralized (deprotonated). Conversely, there are 29 neutralized (protonated) Glu residues out of 60. This suggests that there is a strong asymmetry between neutralizing the two different residues. Additionally, there are twice as many neutralized Glu residues in the protein associated with helix propagation as there are in the N-terminal helix cap. A final observation is that eeEE is prevalent while EEee fails to make an appearance in the top sequences. However, this observation turns out to be a coincidence due to the sharing of common parents in the genetic algorithm.



C)

	Best Sequences	Mean Helicity
1	GEEEEKKKKeeEEKKKKeeeEKKKKGW	0.784128
2	GeEEeKKKKeeEEKKKKeeEEkKKKGW	0.746373
3	GEEEeKKKKeeEEKKKKeeeEKKKKGW	0.764244
4	GeEEeKKKKeeEEKKKKeeEEKKKKGW	0.735334
5	GEeEEKKKKeeEEKKKkeeeEKKKKGW	0.795698

Consensus Sequence

GEEEEKKKKeeEEKKKKeeEEKKKKGW

Figure 3.2: Results from utilizing GADIS with allowing the charge states to float. (A) The residue resolved helicity for the top five charge states discovered by GADIS compared to the model charge state simulation. (B) The cumulative probability of having a given objective score from GADIS. For reference to the degree of improvement the model charge state objective score is shown in red. The farther left the curve is the better the generation of GADIS did. (C) The residue resolved charge states for the top five scoring results from GADIS. The residues are color coded for the charge value where black and lowercase represents neutralized residues. Additionally the consensus charge state is shown.

3.4 Forced Neutralization of Internal Lys Residues Fails to Converge on to Stable Alpha Helices

We attempted to design sequences that favored neutralization of Lys residues by running GADIS using an all neutralized Lys and all charged Glu seed with an additional biasing term for the objective that disfavored charging (protonating) the Lys resides. This approach failed to converge in multiple generations and we interpret this to imply that suggest sequences with neutralized (deprotonated) Lys residues will not form stable helical conformations in this type of sequence, although poly-L-Lysine does form stable alpha helices at high pH values.

A common problem with genetic algorithms is the lack of divergence in the sampling away from the seeds. With the key observations noted above we embarked on a more directed study on the role of different charge permutations. To test directly for the accuracy of the predicted / inferred pattern of neutralization, we tested different patterns such as switching whether the first two Glu residues acids in each block were protonated or deprotonated, followed by the converse. Additionally, we tested specific sequences where we broke up the sequential nature of the neutralizations, such as eEeE. We found that the pattern of neutralization itself was inconsequential. Instead, what was important was the fraction of the Glu residues within each block that was neutralized. This result carries over to the charge states with 25% and 75% neutralization as well, shown in figures 3.3a and 3.3c, respectively. As the fraction of Glu residues that are neutralized increases, so does the thermodynamic stability of the single alpha helical conformations. For 75% and 100% neutralized charge state, the deviations from 100% helicity are driven more from initial condition artifacts, from the simulations starting in a random coil conformation, than from the true thermodynamics of helix folding / unfolding.

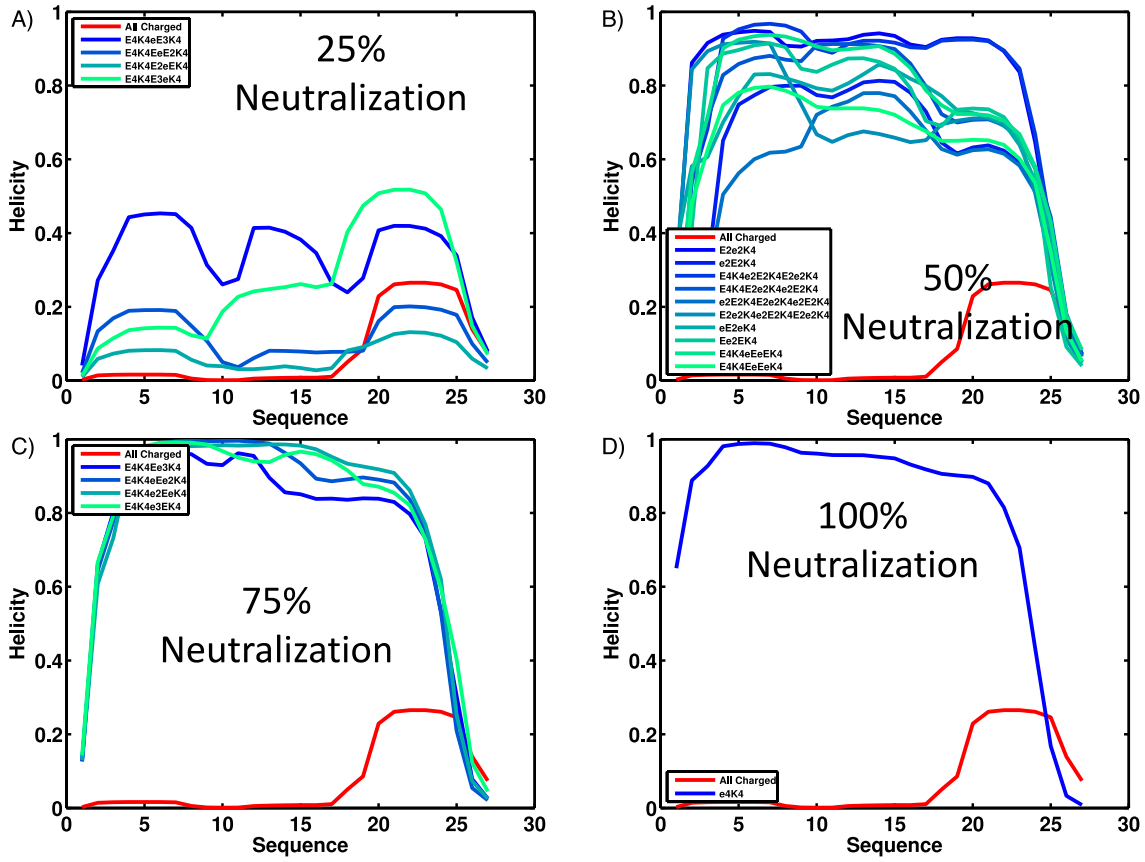


Figure 3.3: Helicity for a series of fractional neutralization and patterns of neutralization. These curves show the residue resolved helicity compared to the model compound charge state in red. The percent neutralization is the percent neutralized in the Glu's not including the first four Glu's which behave uniquely. (A) Multiple patterns of charge states with 25% of the Glu's neutralized. (B) Multiple patterns of charge states with 50% of Glu's neutralized. A collection of these sequences had the four N-terminal Glu's in their neutralized states. (C) Multiple patterns of charge states with 75% of the Glu's neutralized. (D) The 100% neutralized charge state.

3.5 Free energy changes associated with the coupling of charge neutralization and helix formation

To study the thermodynamics in more detail we turned to umbrella sampling with replica exchange. Here, we picked a total of nine representative sequences, four with the N-terminal four Glu residues being charged, four with these neutralized, and the fully charged state, as shown in figures 3.4a and 3.4b, respectively. We calculated the free energy as a function of percent helicity using the multistate Bennett acceptance ratio method[14]. The free energy cost of forming a helix in the fully charged state is an insurmountable being approximately 8 kcal / mol. As the fraction of neutralized Glu residues increases, conformations that are highly helical go from being strongly unfavorable to favorable. There is a significant difference in favorability between the charge states with the first four N-terminal glutamic acids. Keeping these residues charged confers a significant increase in the stability of strongly helical conformations. This is consistent with the increased volume for solvation at the end of the helix as well as interacts with the dipole formed on the axis of an alpha helix. Deprotonated Glu residues are known to stabilize the N-terminal cap of alpha helices[15-17].

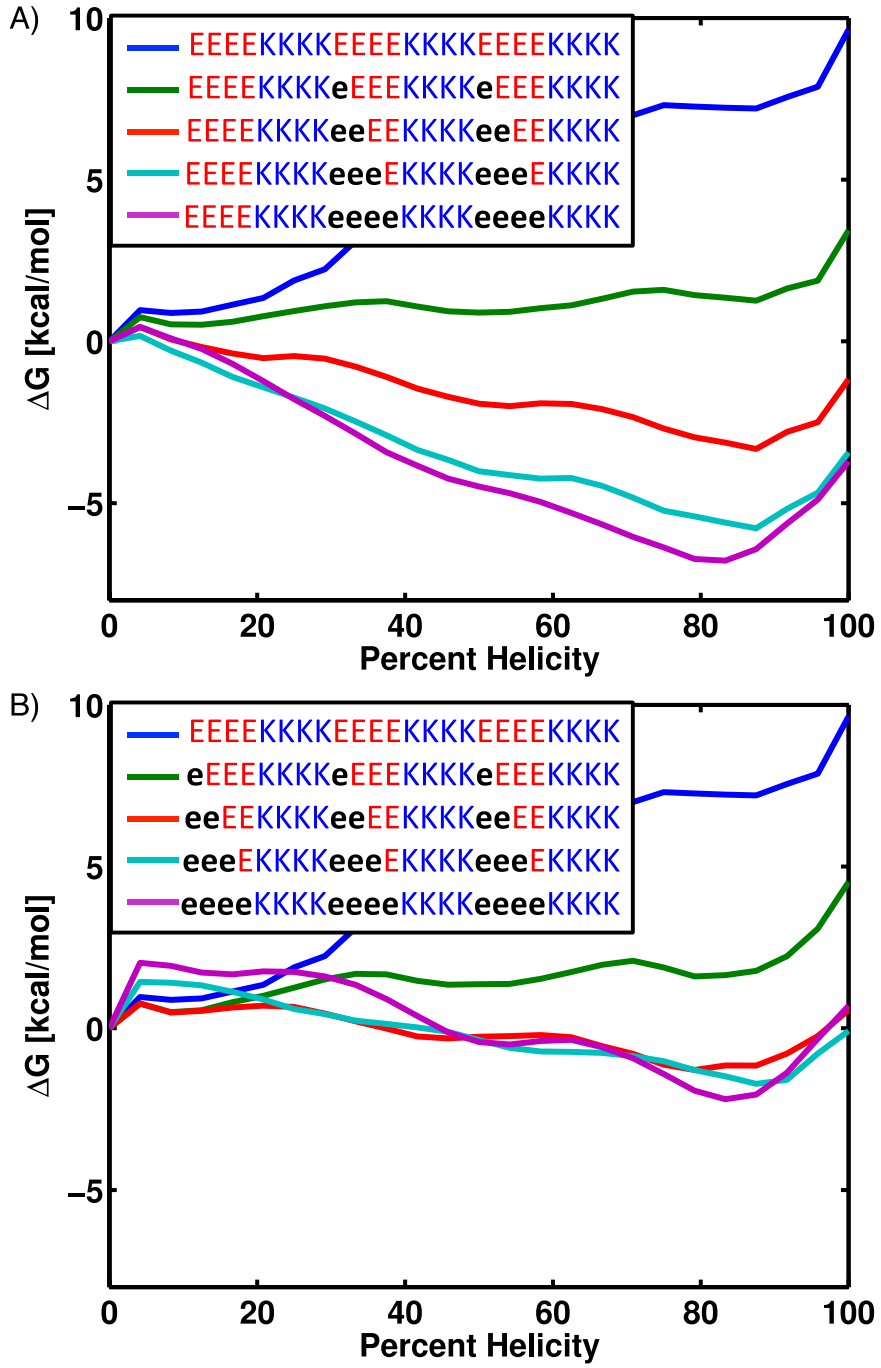


Figure 3.4: Free energy of helix formation for different fractions of neutralization. The free energy of forming a helix seed and propagating it across the entire sequence. Of note, the last 12% helicity is adding the two end glycines and the tryptophan to the single alpha helix. (A) Neutralizing the protein while holding the four N-terminal Glu's in their charged state. (B) Neutralizing the entire protein including the four N-terminal Glu's.

3.6 Main Predictions

Taken together, our simulations yield three main predictions: (1) At neutral pH, E4K4r sequences preferentially a large fraction of internal Glu residues in order to form stable single alpha helices. (2) The four N-terminal Glu residues are an exception in they prefer to be deprotonated in order to initiate helix formation. At low pH, the N-cap should also be neutralized thus reducing the stability of the helix. (4) At high pH, the Glu residues will become deprotonated to and the Lys residues will be preferentially deprotonated. If the destabilizing effects of the former are more pronounced than the helix stabilizing effect of the latter, then we expect a significant diminution of helicity at high pH. These predictions are directly testable using a combination of ultraviolet circular dichroism measured as a function of pH and potentiometric measurements that directly get at the charge state of the E4K4r sequences under different pH conditions.

3.7 Experimental Tests of the Predictions from Simulations

Dr. Ammon Posey, a research scientist in the Pappu lab, performed two sets of experiments to test the predictions summarized above. He measured the pH dependence of the helicity of $(E4K4)_3$ using CD spectroscopy. In figure 3.5a we show representative CD spectra at several different pH values obtained by careful titration with the strong base NaOH, which shows that the peptide is strongly helical over a range of different pH values. In figure 3.5b, we show the molar ellipticities at 208 and 222nm values. These features are associated with helicity. This analysis shows three distinct regions. At low pH, the protein has a modest helicity. At medium pH, the protein is strongly helical. At high pH, the protein appears to be approaching a random coil. These observations are consistent with predictions from the simulations.

Additionally, Dr. Posey performed potentiometric measurements to quantify the pH in solution as a function of titrating the concentration of a strong base in the presence vs. absence of the E4K4r peptide. The current data, which are reproducible, albeit preliminary, are shown in figure 3.6. By fitting the titration curve to what we expect to observe based on model compounds vs. what we obtain if we allow for pK_a shifts, we estimate that ~9 Glu residues have anomalous pK_a values of approximately 9.3. This is to be contrasted with model compound pK_a value of Glu, which are approximately 4.1. These findings are in agreement with the simulation results.

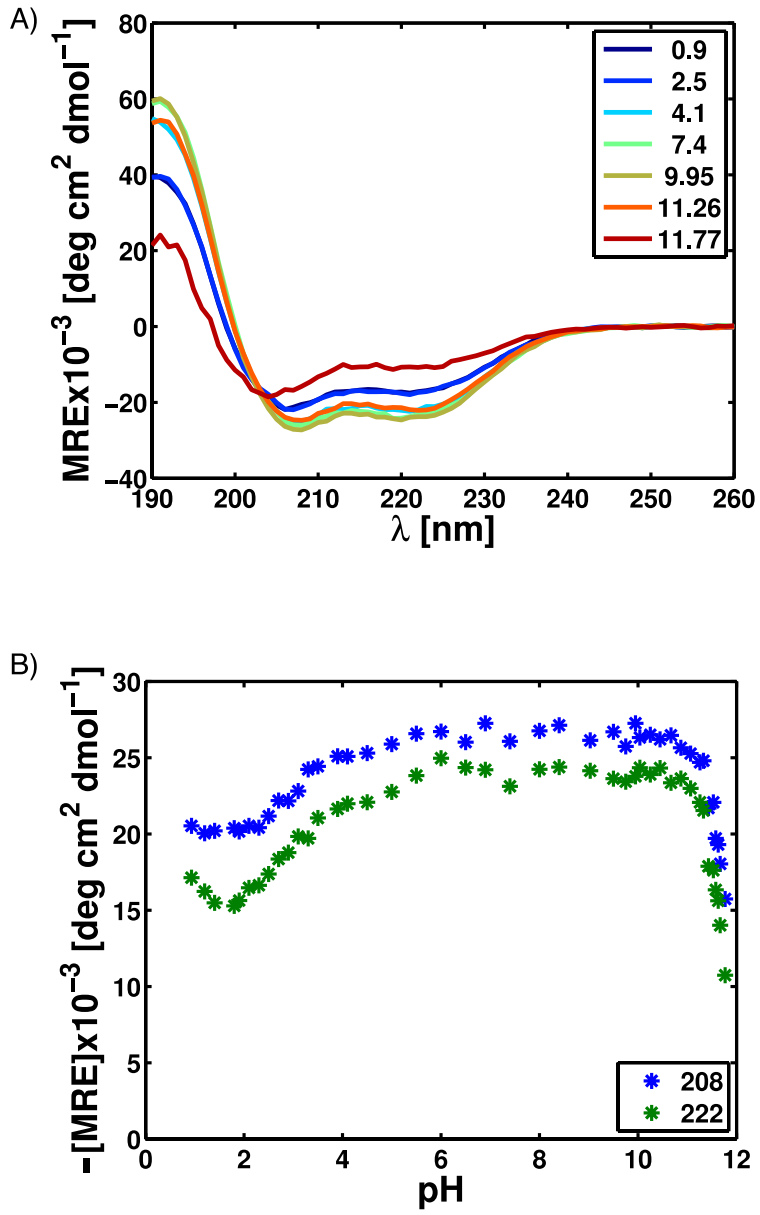


Figure 3.5: UV-CD data for E4K4 peptide as a function of pH. (A) The full spectra for a few representative pH values. (B) A plot of minus the value at 208 and 222nm as a function of pH. This gives an estimate of the degree of helicity and shows that there are three different pH regimes.

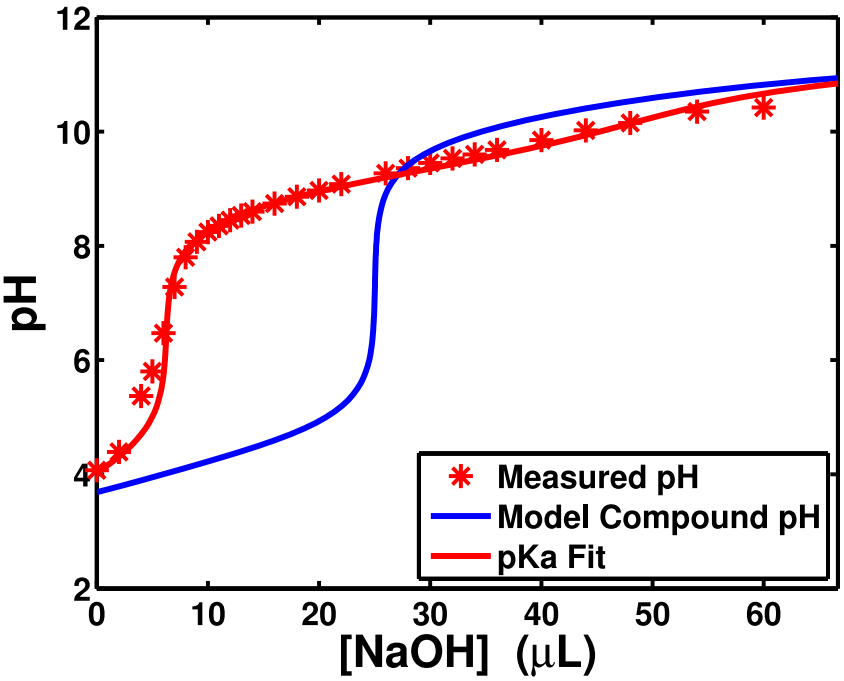


Figure 3.6: Potentiometry for E4K4. This plot shows the expected pH response for E4K4 if it were using model pKa values for the side chains (blue). The measured values are shown in red asterisks and the fit for the experimental data (red).

3.8 Discussion

Based on our simulation results and preliminary experiments we propose that the stability of alpha helical conformations in sequences with blocky patterns of oppositely charged residues is driven by the preferential neutralization of internal acidic groups – Glu residues. Only the N-terminal Glu residues, which form helix-initiating N-caps, remain in their charged states. This implies that instead of instead of a network of intramolecular salt bridges, the stability of alpha helices in sequences with blocky patterns of oppositely charged residues is determined by the preferential neutralization of acidic groups. We find that neutralization of Glu residues are responsible for shifting the observed pKa values significantly above neutral pH creating a native charge state that is far from what the model compound pKa values would suggest. Additionally,

the simulations predict behavior for both low and high pH. At low pH, E4K4r should have a medium degree of helicity as the N-terminal glutamic acids become protonated and simulations show this conformation having a modest stability toward forming helical conformations. At high pH, E4K4 will have a low degree of helicity as the Glu residues revert to their deprotonated states and Lys residues are deprotonated.

We have shown that the experimentally measured pH dependence of helicity follows these trends supporting our simulations results. Our predictions that charge neutralization preferentially favors protonation of Glu residues as opposed to deprotonating Lys residues is being tested using potentiometric measurements. This is currently ongoing but the preliminary data are promising and suggest that at neutral pH $G(E4K4)_3GW$ has about nine neutralized Glu residues, in strong agreement with the simulation results.

The strong deviation between the simulations of this peptide in what we now propose to be the native charge state compared to its model compound prompts an important question that has been largely ignored in the field of IDPs. **Are pKa shifts, aided by sequence contexts, common to various IDP sequences and should these effects be taken into account when making predictions of the degree of disorder and the overall features of IDP ensembles**? To answer this question we need a systematic set of measurements directed at the ordered and disordered states. We also need to perform ABSINTH-based simulations by fixing the chemical potential of the proton, i.e., perform simulations at constant pH. Several constant pH simulation methodologies have been developed to address this question for folded proteins[18-20]. These approaches center on molecular dynamics simulations and in some implementations, one uses periodic Monte Carlo moves that change the degree of protonation of a randomly chosen titratable group. Other approaches based on so-called λ-dynamics treat the degree of protonation

as a scalar that modulates the strength of interactions between the protonating hydrogen with all other atoms and the local geometry of the given sidechain. This appears to work well for proteins that are folded where the native state is prescribed by knowledge of a well-defined three-dimensional structure. However, if there are many unexpected pK_a shifts or the ensembles in one charge state has virtually no overlap with the ensemble in another charge state, then we should expect virtually no convergence in the simulations. This is especially relevant when we think about disorder to order transitions. These transitions have in practice zero overlap. This means that a simulation will stay in some local minimum and never explore the alternative conformations and charge states.

For disordered proteins, we suggest an alternative methodology. A protein with ten charges has 1024 charge states, too many to ever exhaustively sample in simulations. Therefore, instead of passively letting the charge states explore for themselves, as has been developed and utilized in previous studies, we are at the early stages of developing a Markov State Model where we use umbrella sampling simulations to calculate the free energy between ensembles of different charge states. In the long term, we hope to couple this with a machine-learning algorithm that proposes the new states to test in appending to a Markov State Model that is built on the fly.

3.9 Experimental Methods

Synthesis of peptide constructs: Peptides were purchased from Watsonbio Sciences (Houston, TX), at >95% purity with acetylated N-termini and amidated C-termini. Peptides were stored in lyophilized form at –20℃ in sealed containers in the presence of desiccant until use.

Circular Dichroism (CD): A peptide stock solution was prepared from lyophilized peptide in deionized water, and the peptide concentration was determined spectrophotometrically using tryptophan absorbance at 280 nm with an extinction coefficient of $\varepsilon = 5500\ M^{-1}\ cm^{-1}$. CD samples were prepared by diluting the peptide stock solution into various buffered solutions depending on the desired pH, with a final peptide concentration of 15-20 μM. The buffers and the corresponding pH ranges covered are as follows: phosphoric acid, pH 0.9-1.9; formic acid, pH 1.8-2.1; mixtures of formic acid and potassium hydroxide, pH 2.3-4.1; 10 mM sodium phosphate (mixtures of mono- and di-basic) pH 5.0-9.5; mixture of borax and borate, pH 9.0; mixtures of sodium bicarbonate and sodium carbonate, pH 9.95-11.53; sodium carbonate, pH 11.43-11.67. Samples were also taken from an unbuffered titration of 90-95 μM peptide with sodium hydroxide as the titrant, covering the pH range of 3.4-8.4 and measured by CD in a demountable quartz cuvette with a 0.1 mm path length. All other CD samples were prepared in a quartz cuvette with a 1 mm path length. All CD spectra were the average of six scans from 260-190 nm, with a 1 nm step, a 2 second response time, and 50 nm/min scanning speed, using a JASCO J-810 spectropolarimeter. Trends in mean residue ellipticity ([θ], determined as described below) across ranges of pH where overlap in buffering ranges occurred was independent of buffer type. A background CD spectrum was measured for each buffer and subtracted from the peptide CD spectrum. The resulting ellipticity was used to calculate the mean residue ellipticity ([θ], units of deg cm^2/dmol residue) using equation 3.2:

$$[\theta] = \frac{\theta}{(N-1)L_{mm}C_M}\ ; \tag{3.2}$$

In equation (2), θ is the molar ellipticity (mdeg, machine units), N = number of residues, N-1 = number of backbone amides, L_{mm} = path length (mm), and C_M = protein concentration (Molar).

Potentiometry: Peptide was dissolved in pure deionized water at a concentration of about 200 μM and buffer exchanged into degassed pure deionized water using an ultraspin filter with a 3000 kDa molecular weight cut-off. The resulting degassed desalted solution was then diluted into degassed pure deionized water with 100 mM KCl. All water and water/potassium chloride solutions were degassed in sealed vessels under alternating cycles of nitrogen gas and vacuum using a custom-built Schlenk line. Buffers and samples were kept in sealed flasks, and were transferred between sealed containers, as needed, using precision gas-tight μL syringes (Hamilton). Potentiometry measurements were carried out in a custom sample vial consisting of a glass vial sealed with a rubber septum that accommodated the pH probe. Titrant was delivered to the sealed vial through the rubber septum using a precision gas-tight μL syringe fitted with a repeating dispenser (Hamilton). Potentiometric measurements were carried out on an Orion Star A215 potentiometer (Thermo Scientific) using a pHT-micro combination probe with a platinum/Silamid double junction (YSI). A 2 mL sample was titrated with 25-50 mM KOH previously calibrated against potassium phthalate (KHP) prepared in the same manner as the sample. A potentiometric titration was also carried out on "background solvent" (water with 100 mM KCl from the same degassed stock used to prepare sample). Estimates of pK_as were extracted from the potentiometry data using Levenberg-Marquardt non-linear least squares fits to equation 3.3, adapted from de Levie[21, 22]:

$$V_b = V_a \frac{\sum F_a C_a - \Delta}{F_b C_b + \Delta} \tag{3.3}$$

where

$$F_a = \frac{K_a}{K_a + [H^+]} \tag{3.4}$$

and

$$\Delta = [H^+] - [OH^-] = [H^+] - \frac{K_w}{[H^+]} \; ; \tag{3.5}$$

Here, V_b is the volume of titrant added V_a is the sample volume, C_a is the concentration of ionizable groups, C_b is the concentration of titrant, K_a is the acid dissociation constant of the ionizable moiety of interest, and K_w is the dissociation constant of water. Data were also analyzed by the method of Tanford [23-25], where the titration of background solvent was subtracted from the peptide titration in order to determine the number of protons bound per peptide at a given pH.

3.10 Acknowledgements

We are grateful to Rahul Das for stimulating discussions. A grant from the National Science Foundation (MCB1614766) supported this work. TSH is a graduate student scholar of the Center for Biological Systems Engineering at Washington University in St. Louis.

3.11 References

1. Das, R.K. and R.V. Pappu, *Conformations of intrinsically disordered proteins are influenced by linear sequence distributions of oppositely charged residues.* Proc Natl Acad Sci U S A, 2013. **110**(33): p. 13392-7.
2. Mao, A.H., N. Lyle, and R.V. Pappu, *Describing sequence-ensemble relationships for intrinsically disordered proteins.* Biochem J, 2013. **449**(2): p. 307-18.
3. Mao, A.H. and R.V. Pappu, *Crystal lattice properties fully determine short-range interaction parameters for alkali and halide ions.* J Chem Phys, 2012. **137**(6): p. 064104.

4. Sawle, L. and K. Ghosh, *A theoretical method to compute sequence dependent configurational properties in charged polymers and proteins.* J Chem Phys, 2015. **143**(8): p. 085101.

5. Gaspari, Z., et al., *Charged single alpha-helices in proteomes revealed by a consensus prediction approach.* Biochim Biophys Acta, 2012. **1824**(4): p. 637-46.

6. Peckham, M. and P.J. Knight, *When a predicted coiled coil is really a single α-helix, in myosins and other proteins.* Soft Matter, 2009. **5**(13): p. 2493-2503.

7. Swanson, C.J. and S. Sivaramakrishnan, *Harnessing the unique structural properties of isolated alpha-helices.* J Biol Chem, 2014. **289**(37): p. 25460-7.

8. Baboolal, T.G., et al., *The SAH domain extends the functional length of the myosin lever.* Proc Natl Acad Sci U S A, 2009. **106**(52): p. 22193-8.

9. Baker, E.G., et al., *Local and macroscopic electrostatic interactions in single alpha-helices.* Nat Chem Biol, 2015. **11**(3): p. 221-8.

10. Andrews, C.T. and A.H. Elcock, *Molecular dynamics simulations of highly crowded amino acid solutions: comparisons of eight different force field combinations with experiment and with each other.* J Chem Theory Comput, 2013. **9**(10).

11. Debiec, K.T., A.M. Gronenborn, and L.T. Chong, *Evaluating the strength of salt bridges: a comparison of current biomolecular force fields.* J Phys Chem B, 2014. **118**(24): p. 6561-9.

12. Harmon, T.S., et al., *GADIS: Algorithm for designing sequences to achieve target secondary structure profiles of intrinsically disordered proteins.* Protein Eng Des Sel, 2016. **29**(9): p. 339-46.

13. Vitalis, A. and R.V. Pappu, *ABSINTH: a new continuum solvation model for simulations of polypeptides in aqueous solutions.* J Comput Chem, 2009. **30**(5): p. 673-99.

14. Shirts, M.R. and J.D. Chodera, *Statistically optimal analysis of samples from multiple equilibrium states.* J Chem Phys, 2008. **129**(12): p. 124105.

15. Aurora, R. and G.D. Rose, *Helix capping.* Protein Sci, 1998. **7**(1): p. 21-38.

16. Altschuler, E.L., *Alpha helix capping and the conformation of threonine.* Med Hypotheses, 2001. **56**(4): p. 478-9.

17. Aceto, A., et al., *Identification of an N-capping box that affects the alpha 6-helix propensity in glutathione S-transferase superfamily proteins: a role for an invariant aspartic residue.* Biochem J, 1997. **322 (Pt 1)**: p. 229-34.

18. Donnini, S., et al., *Charge-Neutral Constant pH Molecular Dynamics Simulations Using a Parsimonious Proton Buffer.* J Chem Theory Comput, 2016. **12**(3): p. 1040-51.

19. Lee, M.S., F.R. Salsbury, Jr., and C.L. Brooks, 3rd, *Constant-pH molecular dynamics using continuous titration coordinates.* Proteins, 2004. **56**(4): p. 738-52.

20. Mongan, J., D.A. Case, and J.A. McCammon, *Constant pH molecular dynamics in generalized Born implicit solvent.* J Comput Chem, 2004. **25**(16): p. 2038-48.

21. de Levie, R., *Explicit expressions of the general form of the titration curve in terms of concentration: Writing a single closed-form expression for the titration curve for a variety of titrations without using approximations or segmentation.* Journal of Chemical Education, 1993. **70**(3): p. 209.

22. de Levie, R., *General Expressions for Acid−Base Titrations of Arbitrary Mixtures.* Analytical Chemistry, 1996. **68**(4): p. 585-590.

23. Garcia-Moreno, B., et al., *Experimental measurement of the effective dielectric in the hydrophobic core of a protein.* Biophys Chem, 1997. **64**(1-3): p. 211-24.

24. Huang, Y. and D.W. Bolen, *Probes of energy transduction in enzyme catalysis.* Methods Enzymol, 1995. **259**: p. 19-43.

25. Nozaki, Y. and C. Tanford, *[84] Examination of titration behavior.* Methods in Enzymology, 1967. **11**: p. 715-734.

Chapter 4

Disordered Linkers Modulate the Coupling Between Phase Separation and Gelation in Multivalent Proteins

This chapter is adapted from an article under preparation. Tyler S. Harmon and Rohit V. Pappu developed the coarse-grained framework. Tyler S. Harmon performed and analyzed the simulations.

4.1 Introduction

There is growing evidence that a variety of intracellular processes are governed by phase transitions that lead to the formation of biomolecular condensates of protein and RNA molecules [1]. The material properties of biomolecular condensates are consistent with either liquids or gels, depending on the molecular components and cellular state [2-6]. Biomolecular condensates, which are represented by different types of membraneless organelles, are involved in cell signaling [7], transcriptional regulation [8-10], cytoskeletal regulation [6], stress response [11-13], cell division [14, 15], and cytoplasmic branching [16]. Membraneless organelles encompass several proteins, RNA molecules, and metabolites [5, 17]. The protein components of membraneless organelles are classified as scaffolds and clients [18]. Scaffold proteins drive phase transitions that give rise to membraneless organelles, whereas client molecules preferentially partition into these bodies [18, 19]. The scaffold proteins have distinct features, the most prominent being multivalency of either well-folded protein domains or short linear motifs

that are encompassed in low complexity disordered regions [1, 6, 20]. In fact, multivalency is a defining hallmark of many proteins that control cell signaling where they serve as scaffolds or adaptor proteins [21-36]. The concept of valence refers to the number of interaction domains within a multivalent protein. The ligands of multivalent proteins can be other multivalent proteins or polynucleotides with a multiplicity of interaction motifs.

The simplest linear multivalent proteins involve multiple protein-protein or protein-nucleic acid interaction domains connected by intrinsically disordered linkers (Fig 4.1a). Mean field polymer theories predict that the multiplicity of complementary interactions amongst multivalent proteins and their multivalent ligands will give rise to so-called sol-gel transitions [37-39]. Of direct relevance to cell signaling are *chemo-reversible* transitions that are controlled by the bulk concentrations (c_b), *i.e.*, chemical potentials, of interaction domains and their ligands [1, 6, 7, 18, 40, 41]. Sol-gel transitions are characterized by the existence of a concentration threshold known as the gel point or c_g [37-39]. If c_b is smaller than c_g, the multivalent proteins and their multivalent ligands form a dispersed phase or sol of largely unbound molecules. Beyond the gel point c_g, a majority of the multivalent proteins and their multivalent ligands are incorporated into large, system-spanning networks known as gels that are stabilized by physical crosslinks [7, 33]. This refers to non-covalent interactions between specific binding partners or motifs. A gel can be a dilute or dense liquid, an amorphous, crystalline, or semi-crystalline solid, or one of many liquid-crystalline phases [42]. The different types of gels share the common feature of being reached from their corresponding sol phases via a change in the connectivity within the system.

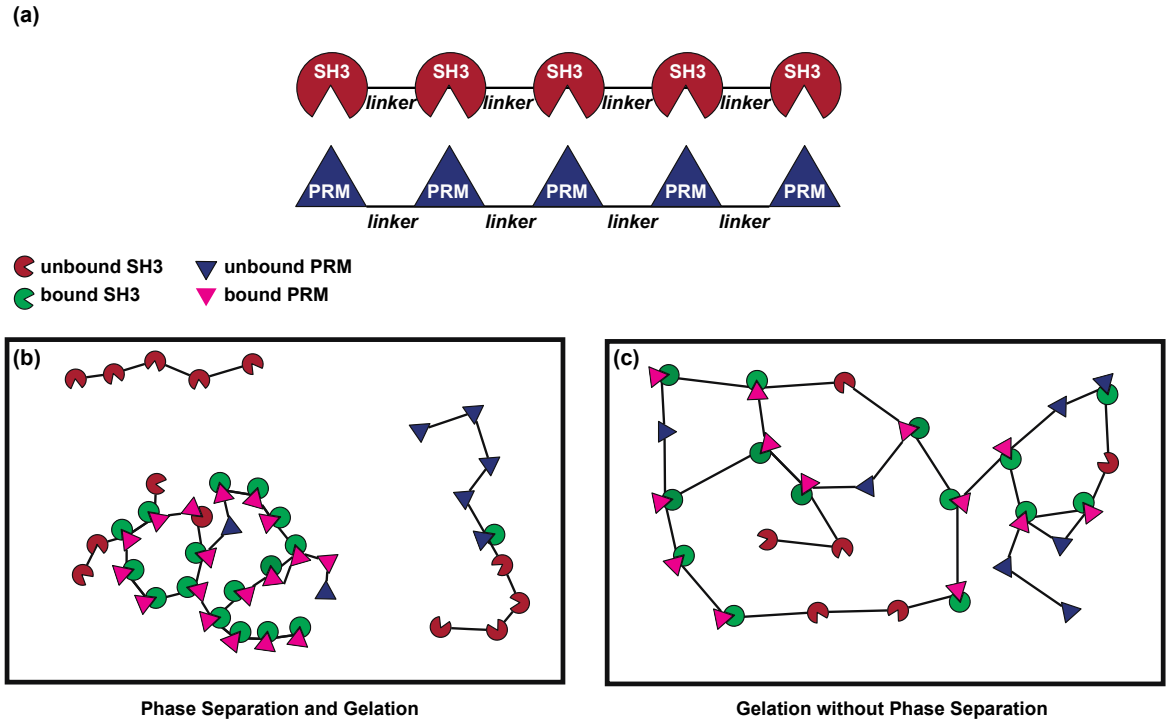


Figure 4.1: Phase separation and gelation are distinct types of transitions. (a) Schematic of a synthetic multivalent system that undergoes liquid-liquid phase separation and gelation. SH3 domains bind to proline-rich modules (PRMs). Multivalent SH3 and PRM proteins result from the tethering of multiple SH3 domains (or PRMs) by flexible linkers. (b) Schematic of phase separation and gelation in a system of pentavalent SH3 and PRM proteins. Cooperative binding of multivalent proteins drives phase separation: If the bulk concentration of interaction domains is higher than a threshold concentration c_s, a dense phase comprising of multivalent SH3 and PRM proteins will be in equilibrium with a dispersed phase of unbound proteins. The dense phase is typically a spherical droplet and within the dense phase a droplet-spanning network can form. The latter is a gelation transition whereby a majority of the molecules within the droplet become part of a single connected cluster. The bounding box depicts the volume occupied by molecules in the dispersed phase. (c) A sol-gel transition can also occur without phase separation. If the bulk concentration of interaction domains is higher than a threshold value c_g (typically, $c_g > c_s$), then a system-spanning network can form across the entire volume. In this scenario, a majority of the multivalent proteins become part of the network, and this transition is realized without any change in density.

Many polymeric systems undergo phase separation [43, 44]. These phase changes are a different archetype of chemo-reversible transitions. Above a bulk concentration threshold, designated as c_s, the polymer solution separates into a dense polymer-rich phase that coexists with a dilute, polymer-deficient phase. The dense phase could be a liquid, a solid, or a liquid crystal. The precise state is determined by the extent and directionality of long-range ordering that accompanies the change in density. Liquid-liquid phase separation refers to a change in

density whereby a dilute liquid, deficient in polymers, coexists with a dense liquid that is rich in polymers [43, 44].

Sol-gel transitions arise due to changes in connectivity, which refers to a change in the extent of crosslinking within a system of multivalent molecules. In contrast, phase separation refers to the change in the density of multivalent molecules. The two transitions can be coupled to one another whereby phase separation promotes gelation because the concentration of interaction domains within the dense liquid is above the gel point, c_g (Fig. 4.1b). Sol-gel transitions can also be decoupled from phase separation, as shown in Figure 4.1c.

Semenov and Rubinstein developed a mean field model for the degree of coupling between liquid-liquid phase separation (referred to hereafter as phase separation) and sol-gel transitions [45, 46]. They considered linear, flexible polymers of associative "stickers". The stickers in their model are akin to binding domains in multivalent proteins. Pairs of stickers associate with a binding energy that is a multiple of the thermal energy k_BT. Here, k_B is Boltzmann's constant and T is the temperature. The linkers between stickers do not associate with the stickers. All auxiliary interactions involving the linkers are quantified in terms of an excluded volume parameter, v_{ex}. This parameter quantifies the average volume that is excluded by linker residues for interactions with the solvent [47].

The sign and magnitude of v_{ex} are determined by the balance of linker-linker, solvent-solvent, and linker-solvent interactions [47]. In a poor solvent, v_{ex} is negative because linker-linker interactions are preferred to linker-solvent interactions. In this scenario, the linkers encode an additional driving force for phase separation, and the strength of these interactions can either strongly couple phase separation to gelation. In a good solvent, v_{ex} is positive because linker-solvent interactions are preferred. Finally, in a theta solvent, the linker-linker and linker-solvent

interactions are counterbalanced such that $v_{ex} \approx 0$. While the phase behavior for negative values of v_{ex} is intuitive, Semenov and Rubinstein predicted a surprising result that the formation of a reversible network *viz*., a physically cross-linked gel is always accompanied by phase separation, even for so-called marginal solvents where v_{ex} is either close to zero or slightly positive. However, phase separation is suppressed, when v_{ex} becomes increasingly positive. These results suggest that the sequence-encoded excluded volume of disordered linkers should make a direct contribution to the phase transitions of multivalent proteins [48]. Indeed, a recent study highlighted the role of the highly conserved L1 linker within the adaptor protein Nck, a poly-SH3 protein, as a modulator of the driving force for phase separation and gelation [41]. In this study, the L1 linker was found to house an auxiliary linear motif, involving a stretch of basic residues. This contributes additional contacts through complimentary electrostatic interactions between the basic stretch on the L1 linker and the acidic surface of SH3 domains.

The key question is if the excluded volume of linkers, irrespective of whether or not they house specific auxiliary motifs, can modulate the phase transitions of multivalent proteins, specifically the coupling between phase separation and gelation. Answering this question has direct relevance for designing phase diagrams by modulating the sign and magnitude of v_{ex} that is encoded by disordered linkers. It is also relevant for understanding how phase transitions might be tuned *in vivo* through post-translational modifications of linkers or competing interactions among multivalent proteins with similar interaction domains albeit different linkers.

Recent studies have shown that disordered regions of proteins display a rich encoding of sequence-to-conformation relationships [48]. The sign and magnitude of v_{ex} is not always zero as would be expected for generic random coils. Instead, the charge content and the patterning of oppositely charged residues directly impact the sign and magnitude of v_{ex}. Importantly,

bioinformatics analysis of multivalent proteins shows a wide range of possibilities for the sequences of linkers in these systems (Fig. 4.2).

In this work, we have designed and deployed a coarse-grained lattice model for multivalent proteins and their multivalent ligands to assess the synergy between valence and the excluded volume of linkers on the coupling between phase separation and gelation. Our numerical results reproduce findings from the experiments of Li et al. [6] for synthetic poly-SH3 and poly-PRM systems and also recapitulate the predictions of Semenov and Rubinstein. This result demonstrates that the theoretical predictions hold for finite-sized, biologically relevant multivalent proteins even without the simplifying assumptions of mean field theories.

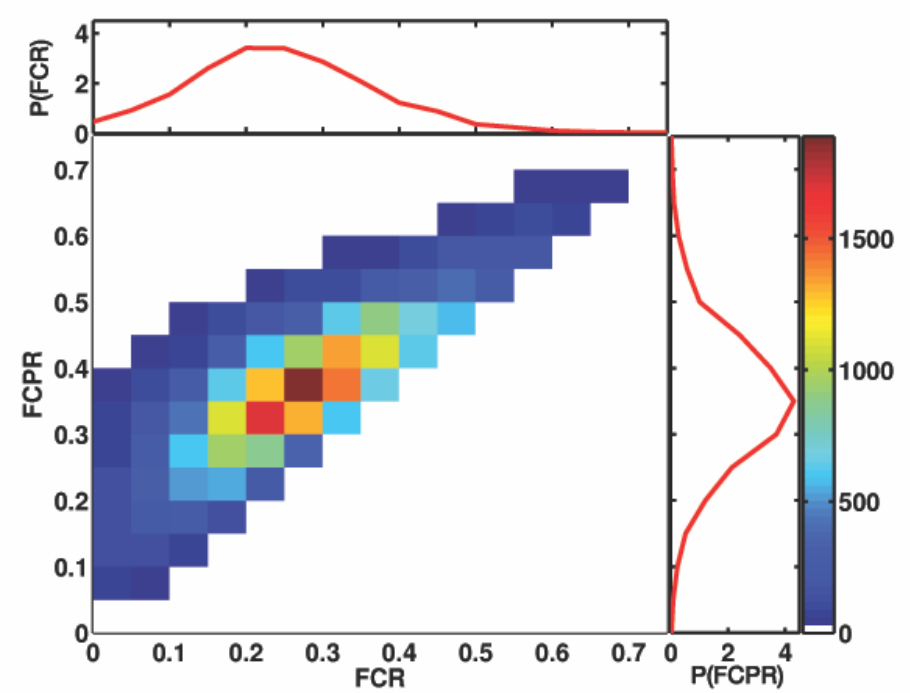


Figure 4.2: Genomic analysis of disordered regions from human proteome. We sorted every region of greater than 20 residues in the human protein that is predicted to be disordered. The regions are sorted based on the fraction of the sequence that has a preference for solvation. On the x-axis we have the fraction of residues that are charged, which under the majority of conditions have a strong preference for being well solvated. On the y-axis we have the fraction of residues that are either charged or a proline. Under certain conditions proline has a strong preference for either expanding the conformation or remaining well solvated. On the top and right are the respective 1-D histograms for each dimension. These histograms show that most disordered regions should be expected to have behaviors ranging from Flory Random Coils to Self Avoiding Random Coil.

4.2 Design of Lattice Simulation

Our focus is on flexible, intrinsically disordered linkers in multivalent systems and their contributions to the driving forces for phase transitions. The goal is to compute full phase diagrams for a given system of multivalent proteins. These calculations require that we include hundreds of multivalent proteins in the simulation setup and perform a series of simulations at different concentrations in order to uncover the concentration-dependent phase diagrams. Such calculations are computationally intractable with traditional approaches based on all atom molecular dynamics or Monte Carlo simulations. Coarse-grained lattice models have played a seminal role in uncovering key concepts in the physics of phase transitions. Moreover, phase transitions arise from the collective synergies among a small number of degrees of freedom that govern the key order parameters. Hence, coarse-graining is inherent to the phenomenon of phase transitions.

We took advantage of this intrinsic feature of phase transitions to develop and deploy coarse-grained lattice models to mimic the synthetic system of poly-SH3 and poly-PRM polymers (Fig. 4.1a). In this model, each interaction domain, *viz*., SH3 or PRM is modeled as a single bead with excluded volume corresponding to a single lattice site. SH3 domains and PRMs can bind to one another and form a 1:1 complex. Flexible linkers connect interaction domains within each multivalent protein. The linker length is cast in terms of the number of lattice sites. Our focus is on coil-like linkers for which $v_{ex} \geq 0$. Accordingly, we consider two limiting cases *viz*., the Flory random coil (FRC) linkers and the self-avoiding random coil (SARC) linkers. For the FRC linkers, $v_{ex} = 0$ and this scenario is captured using so-called implicit linkers (Fig. 4.3a). The FRC linkers are modeled by imposing an infinite square well potential to ensure that the lattice spacing between tethered interaction domains does not exceed *n,* which is the linker length in

terms of the number of lattice sites. For the SARC linkers with positive excluded volume, we use so-called explicit linkers as shown in figure 4.3b. For a SARC linker of length n, we use n non-interacting beads, where each bead is constrained to occupy adjacent vertices on the lattice. Each explicitly modeled linker bead has a finite excluded volume corresponding to one lattice site and linkers of length n require n–1 beads. Atomistic simulations showed that the dimensions of coil-like disordered sequences tethered to SH3 domains are such that each linker bead corresponds to 7-8 residues [49]. Therefore, the linker length can be estimated in terms of number of residues for coil-like linkers to be: $N_r \approx 7n$.

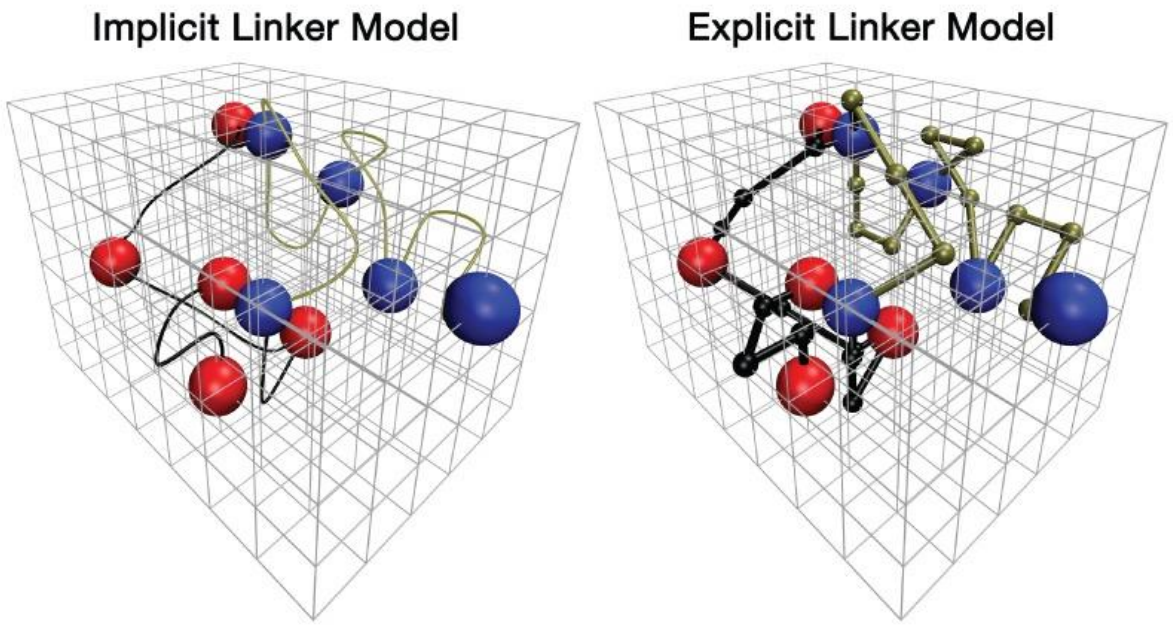


Figure 4.3: Coarse-grained bead-tether lattice models for modeling the phase behavior of multivalent proteins. All simulations were performed using 3-dimensional cubic lattice models. In these models, poly-SH3 and poly-PRM proteins were modeled as bead-tether polymers where the red beads mimic an SH3 domain, the blue beads mimic PRMs, and the black tethers mimic linkers that connect domains / modules to one another. A lattice site was assigned to each of the SH3 domains and PRMs. We used two limiting models for the linkers. The figure on the left shows an implicit linker model. To mimic FRC linkers, each linker was modeled using a distance restraint between tethered domains and the linker itself does not occupy any lattice sites. SARC linkers were mimicked using an explicit linker model, where non-interacting beads corresponding to a prescribed number of lattice sites were assigned to each linker based on the linker length.

4.3 Order Parameter for Gelation

We adopt Flory's definitions and refer to gelation as a connectivity transition whereby the multivalent proteins become part of a connected network that spans the entire system. A system-spanning network incorporates most of the molecules into a single connected cluster. We analyzed each configuration of multivalent proteins to detect the formation of connected clusters. Within every configuration, each interaction between bound SH3 domains and PRMs i.e., *nodes* is referred to as an *edge*. Additionally, linkers that tether domains together are also counted as edges. The connected cluster with the largest number of nodes is designated as the single largest cluster and this quantity, calculated across the entire equilibrium ensemble of configurations, yields our estimate of ϕ_c.

4.4 Order Parameter for Phase Separation

Unlike gelation, phase separation results from a change in density within the system. Accordingly, the spatial dimensions occupied by all multivalent proteins on the lattice serves as a useful proxy for detecting changes in density within the system. We calculate these dimensions in terms of the radii of gyration over all proteins. To calibrate this number, we compute a ratio ρ, which is defined as: $\rho = \left(\frac{R_g^{\text{lattice}}}{R_g^{\text{proteins}}} \right)$. Here, the numerator is the radius of gyration of the entire lattice and the denominator is the ensemble-averaged radius of gyration over all the proteins in the system. The parameter ρ quantifies the relative spatial dimensions of all multivalent proteins and it is directly related to the relative density of the proteins. If ρ is unity, then the proteins are uniformly dispersed through the lattice. Conversely, if ρ increases beyond unity, then the system

has undergone a density transition whereby the proteins occupy fewer numbers of lattice sites than are available to them.

4.5 Sol-gel Transitions and Phase Transitions are Strongly Coupled for Multivalent Proteins with FRC Linkers

We performed a series of Monte Carlo simulations using the coarse-grained lattice model for poly-SH3 and poly-PRM systems of valence 3, 5, and 7 and all combinations of these valencies. The linker length *n* was set to five in all the simulations. Panels (a) and (b) figure 4.4 show the evolution of ϕ_c and ρ, respectively for systems with FRC linkers. Each sub-plot in figure 4.4a shows the value of ϕ_c for a particular combination of PRM and SH3 domain valence and these values are shown as a function of the concentrations SH3 domains and PRMs. In accord with the results of Li et al., figure 4.4a establishes two distinctive features of multivalent systems: For a given combination of SH3 and PRM valencies, we observe a sharp increase in the values of ϕ_c as the concentrations of SH3 domains and PRMs increase. This behavior is consistent with the expected features of a sol-gel transition. Secondly, as valencies increase, the concentrations of SH3 domains and PRMs at which ϕ_c increases sharply becomes smaller. This recapitulates predictions from mean field theories for sol-gel transitions [37, 39, 42] and the experimental observations of Li et al. [6] who established the importance of multivalency as a driver of phase transitions in signaling molecules. Figure 4.4b shows the evolution of ρ for each of the multivalent systems. Systems displaying the sharpest transitions in terms ϕ_c are also accompanied by sharp increases in the values of ρ. This is illustrated in the plots for the 7:5, 7:7, 5:5, and 5:7 systems. Here, the x:y designation refers to the valence of SH3 domains : the valence of PRMs. In contrast, the 5:3, 3:3, and 3:5 systems show gelation transitions, albeit at

considerably higher concentrations of modules with negligible changes to ρ. In each simulation, the initial conditions correspond to the multivalent proteins being randomly dispersed across the cubic lattice. A representative post-equilibration configuration for a 7:7 system with FRC linkers of length five is shown in figure 4.4e for a value of ϕ_c being well above the gel point. The bounding box corresponds to the volume of the simulation cell and provides a perspective regarding the change in density within the system. A dense (high ρ) spherical droplet forms that is in equilibrium with a small number of dispersed proteins. The poly-SH3 (red molecules) and the poly-PRM (blue molecules) are well mixed within the droplet. These observations highlight valence-dependent coupling between gelation and phase separation that we observe for the system with FRC, zero excluded volume, linkers.

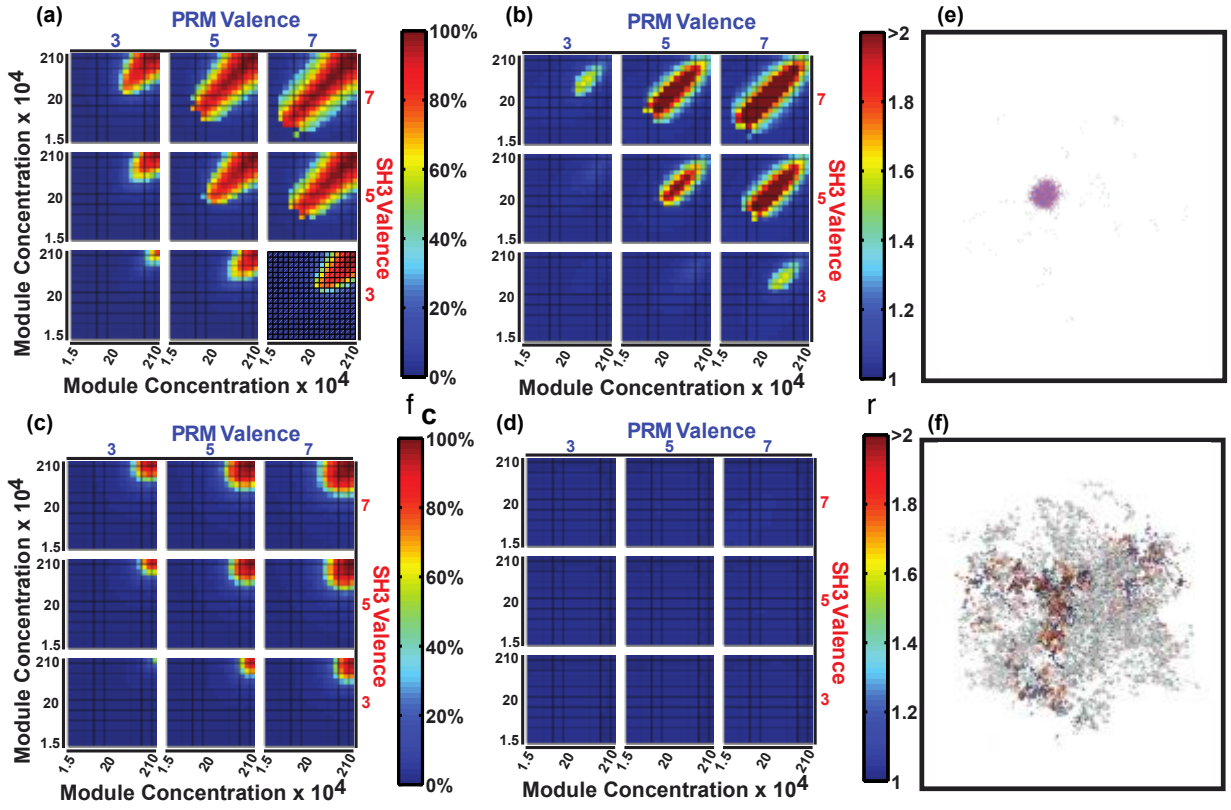


Figure 4.4: Comparative phase behavior of multivalent proteins with FRC linkers versus SARC linkers. (a) A 3×3 matrix of plots that quantifies the fraction, ϕ_c, of SH3 and PRM proteins in the single largest cluster. In this system, FRC linkers connect domains within multivalent proteins. Within each subplot, a heat map quantifies the value of ϕ_c as a function of the concentration of PRMs along the abscissa and the SH3 domains along the ordinate. (b) For the system studied in panel (a), the 3×3 matrix of plots quantifies the ratio ρ of the radius of gyration of the lattice to that of all of the proteins in the system. Within each subplot, a heat map quantifies the value of ρ as a function of the concentration of PRMs along the abscissa and the SH3 domains along the ordinate. Panels (c) and (d) are equivalent plots of panels (a) and (b), respectively for the system where SARC linkers connect domains within multivalent proteins. In panels (a) – (d), the module concentrations are in units of number of modules per lattice voxel. Panels (e) and (f) show representative snapshots obtained for the two limiting systems, *viz*., multivalent proteins connected by FRC linkers – panel (e) – versus SARC linkers – panel (f). Panel (e) shows a dense droplet coexisting with dispersed molecules, whereas panel (f) shows a system-spanning network of molecules forming without an accompanying change in density.

4.6 Sol-gel Transitions are Weakened and Phase Transitions are Strongly Suppressed for Multivalent Proteins with SARC Linkers

Figures 4.4c and 4.4d summarize the equivalent results obtained for poly-SH3 and poly-PRM systems with SARC linkers. Here, five excluded volume beads were modeled explicitly for each of the linkers between SH3 domains and PRMs. The results provide a striking contrast to those for FRC linkers. All systems, except the 3:3 system, show a sharp increase in ϕ_c past a threshold SH3 / PRM concentration. The concentrations at which the transitions are realized are at least an order of magnitude higher than those observed for the systems with FRC linkers. Additionally, none of the systems show any discernible changes to ρ. This implies that sol-gel transitions are realized only when the concentrations are large enough to enable networking through random encounters. The positive excluded volumes of SARC linkers strongly suppress phase separation and gelation requires considerably higher concentrations when compared to multivalent proteins with FRC linkers. Figure 4.4f shows how a sol-gel transition *i.e.*, a system-spanning network forms in the absence of phase separation. Our observations are congruent with the predictions of Semenov and Rubinstein [45]. The implication is that the theoretical underpinnings are applicable even to finite-sized systems with correlated fluctuations as opposed to just infinitely long chains described using mean field theories.

4.7 Calculation of the Gel Point

The heat maps in panels (a) and (c) of Figure 4.4 provide a visual assessment of the sharp increase in ϕ_c as a function of SH3 domain / PRM concentrations. To enable quantitative

comparisons across different systems, it is necessary that we calculate the critical value of ϕ_c, which we is designate as ϕ_{cc} and use as our numerical proxy for the gel point. Figure 4.5 summarizes our approach for computing ϕ_{cc} for a system with prescribed values for the valence, V as well as the binding energy between interaction domains *viz*., SH3 domains and PRMs. We performed simulations of random percolation models, without accounting for linkers or the structure of the lattice models. Each simulation takes the valence, the number of multivalent proteins, and the fraction of bound modules as inputs. The value of ϕ_c is calculated for each prescribed value of the fraction of bound modules and these are shown as solid sigmoidal curves in figure 4.5. The theories of Flory [37, 38] and Stockmayer [39] can be used to calculate ϕ_{cc} analytically for given values of V and the binding energies, as detailed in the methods section. These are shown are vertical dashed lines in figure 4.5. For a given valence V, the horizontal intercept that passes through intersection of the vertical dashed lines and the solid curve defines the value of ϕ_{cc}, which turns out to be ≈ 0.17. The concentration of modules at which ϕ_c becomes greater than 0.17 is taken to be the value of the gel point, c_g for the system of interest. We can therefore calculate the value of c_g directly from our simulations for the multivalent proteins and compare this to the value of c_g that is estimated from Flory-Stockmayer theories. Linkers do not make any contributions to the structure of Flory-Stockmayer theories and the value of c_g provides an important touchstone for quantifying the influence of linkers on phase transitions.

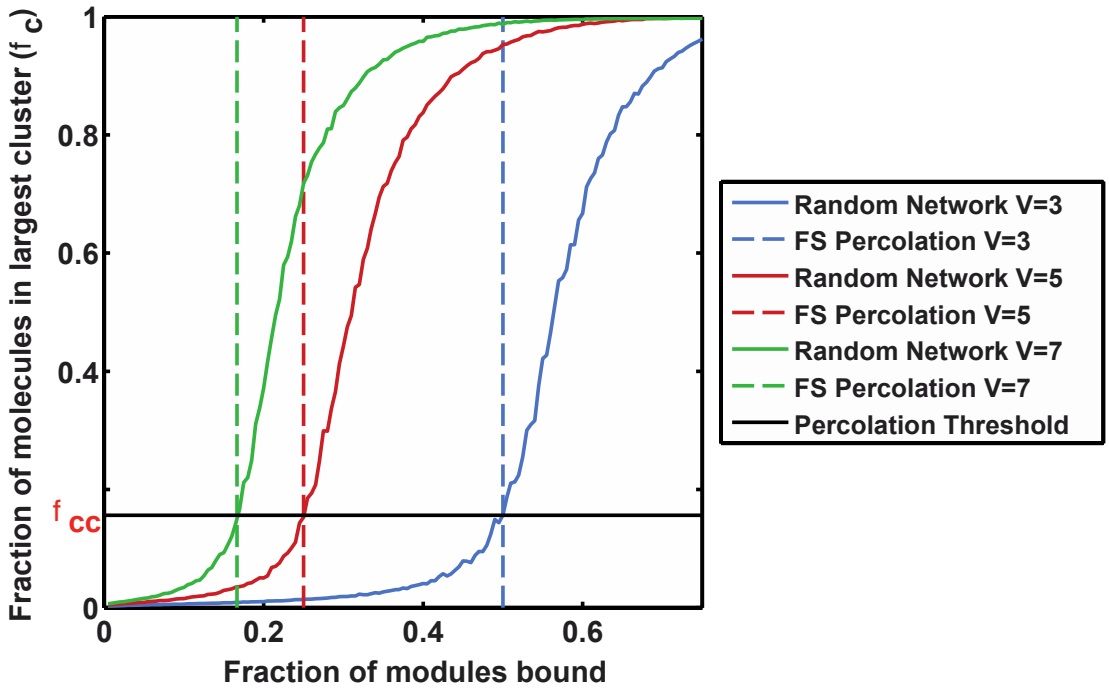


Figure 4.5: Estimating ϕ_{cc}. The fraction of molecules in the largest cluster, ϕ_c, is plotted against the fraction of SH3 domains and PRMs that are bound. ϕ_c was calculated using a random network model (see methods) and for a prescribed affinity between interaction domains. ϕ_c shows a sigmoidal transition that shifts to the right for systems of lower valence (V). For each system, the dashed vertical lines quantify the percolation thresholds, which refer to the fraction of modules for a given valence V that must be bound in order to make a percolated network as prescribed by the theories of Flory and Stockmayer. For a given system of multivalent proteins, the intersection between the solid sigmoidal curve and the dashed vertical line quantifies the value of ϕ_{cc}.

4.8 A Dimensionless Parameter Quantify the Coupling Between Gelation and Phase Separation

Simulations afford the advantage of quantifying the evolution of two distinct order parameters *viz*., ϕ_c and ρ. This allows us to quantify the changes in connectivity and density and infer the presence or absence of coupling between sol-gel transitions and phase separation. Here, we introduce a dimensionless parameter, c^*, which is the ratio of the actual value of c_g to the value

of c_g that is obtained from Flory-Stockmayer theories. We designate the latter as $c_{g,FS}$ and define c^* as: $c^* = \left(\frac{c_g}{c_{g,FS}} \right)$. The value of c^* can be less than, equal to, or greater than one.

Figure 4.6a shows a plot of c^* as a function of linker lengths for 3:3, 5:5, and 7:7 systems with FRC linkers. For long linker lengths, $n > 15$, c^* converges to one. This recovery of the Flory-Stockmayer limit is reasonable since the domains should interact independently, without cooperativity, when the FRC linkers are sufficiently long. Binding domains interact independently with one another in the Flory-Stockmayer theory. In the short linker limit, $n \leq 2$, the value of c^* is greater than one. These linkers are too short and network-terminating oligomers of poly-SH3 and poly-PRM proteins become dominant. Interestingly, for linker lengths in the range $3 \leq n < 12$, we observe that the value of c^* is less than one, and the lowest values of c^* are realized for linkers of length 3-4. FRC linkers within a defined length range engender positive cooperativity by drawing domains together and increasing the apparent affinities through an avidity effect. For linker lengths in the optimal range, the degree of positive cooperativity increases with increasing valence. This effect weakens with increasing linker lengths, because of increasing surface tension that comes from having to sacrifice SH3 domains and PRMs to the surface of the dense droplet. When this energy penalty counterbalances the gain in apparent affinity that is mediated by the linkers, the positive cooperativity is lost and the favorable translational entropy of the modules in the long linker limit is characterized by c^* values of unity.

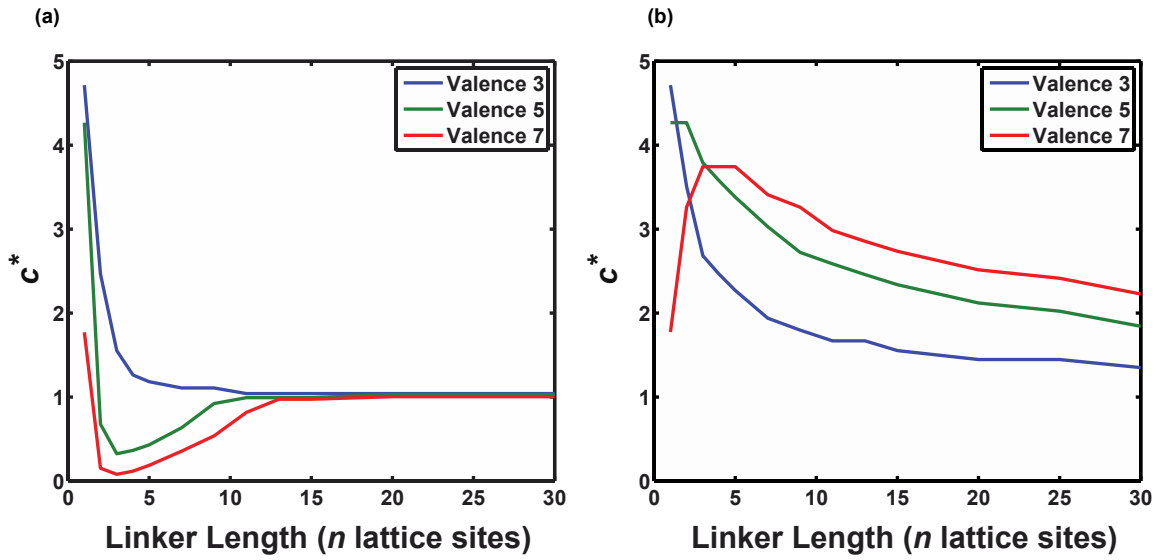


Figure 4.6: Degree of coupling between phase separation and gelation. (a) Plot of c^* as a function of linker length for three symmetric multivalent systems connected by FRC linkers. There is an optimal range for linker lengths where $c^* < 1$, implying a positive cooperativity that gives rise to phase separation and gelation within droplets. For long linkers, c^* converges to unity, implying an absence of cooperativity and the absence of phase separation, in accord with the Flory-Stockmayer theory. (b) Plot of c^* as a function of linker length for three symmetric multivalent systems connected by SARC linkers. The value of c^* is greater than unity for all linker lengths. This points to the suppression of phase separation by linkers with high excluded volume, and a shifting of the gel point to higher concentrations vis-à-vis the threshold predicted by Flory-Stockmayer theory.

Figure 4.6b shows a plot of c^* as a function of linker lengths for 3:3, 5:5, and 7:7 systems with SARC linkers. Unlike the behavior of the FRC linkers, c^* is greater than unity for all the linker lengths we studied and this is true irrespective of the linker lengths. The convergence to unity, even for long linker lengths is compromised, and this suggests a dominance of the excluded volume, which points to preferential solvation of the linkers, that inhibits phase transitions in general, and completely suppresses phase separation. The presence of explicit linkers lowers the apparent affinity through negative cooperativity because the linkers, with their finite excluded volume, can inhibit productive associations between domains. This becomes less of an issue as the linkers become longer. If one corrects intrinsic affinity to account for the weakened apparent affinity, then the convergence of the systems with long linkers to the Flory-

Stockmayer limit is recovered (Fig. 4.7). However, the profiles do not change qualitatively and this points to fundamental differences between systems with FRC versus SARC linkers.

The analysis in this section introduced a dimensionless parameter that provides a measure of cooperativity in the system and a measure of the coupling between phase separation and gelation. If $c^* = 1$, then the interactions between domains are independent of one another and the multivalent proteins undergo sol-gel transitions without phase separation as in the limit of Flory-Stockmayer theories. Conversely, if $c^* < 1$, then the linkers engender positive cooperativity whereby the apparent affinity of the domains for one another becomes higher than the intrinsic affinity due to an avidity effect. This enables a cooperative transition, which leads to a sharp change in density, thus lowering the concentration at which the critical, system-spanning clusters are formed. Finally, if $c^* > 1$, then the linkers engender negative cooperativity because the preferential solvation of linkers crowds out the binding domains. Accordingly, the positive excluded volume of linkers suppresses phase separation and weakens the driving forces for gelation.

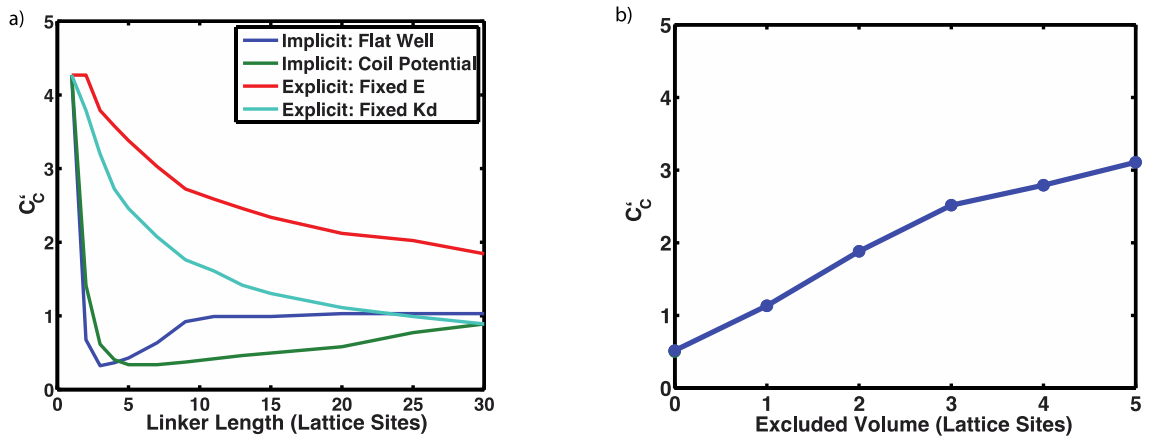


Figure 4.7: Critical Concentrations while correcting for two body terms. (a) The critical concentrations as a function of linker length for four different models. Implicit: Flat Well (blue) and Explicit: Fixed E (red) are the models that have been described previously. The Implicit: Coil Potential (green) and Explicit: Fixed Kd (cyan) models correct for the 2-body differences between the two types of linkers. The Implicit: Coil Potential has a modified end-to-end distance for the linkers so this implicit linker model has end-to-end distances that match the those that are sampled with the explicit linker. The Explicit: Fixed Kd has a modified interaction strength to compensate for the local excluded volume associated with having a excluded volume linkers attached to a domain. If the difference in phase behavior for the two linkers could be explained by their differences in local effects then we would expect the cyan and green curves to be close to overlapping. These corrections do little to qualitatively fix the differences between the models.

4.9 Modulating the Coupling Between Phase Separation and Gelation

In the preceding discussion we focused on the limiting scenarios afforded by the choice of FRC versus SARC linkers. The interaction strengths between domains can be modulated by systematic mutations. Similarly, the magnitude of sequence-encoded excluded volumes can be altered by redesigning the linkers or through posttranslational modifications. Accordingly, we asked if the coupling of phase separation and gelation could be modulated by changes to the intrinsic affinities and to the excluded volumes of linkers. To answer this question, we focused on the 5:5 system with a linker length of $n = 5$. We quantified how the coupling between phase separation and gelation changes as a function of the interaction affinity between the SH3

domains and PRMs and linker excluded volumes. The latter was titrated by fixing the linker length and prescribing the number of linker beads that were modeled implicitly (FRC limit) versus explicitly (SARC limit). The envelope of the two-phase regime is delineated by quantifying the value concentration range where the value of ρ becomes greater than 1.08. This value was chosen by visual inspection of the simulated configurations. We tested the robustness of our findings by changing the threshold value of ρ. The delineation of the two-phase regime becomes sensitive to the choice of the threshold value of ρ near the critical point where the inferred coexistence curve terminates. However, away from the critical point, the delineation of the two-phase regime is insensitive to the choice of the threshold value for ρ.

Figure 4.8a shows the full phase diagram that we compute using a combination of Flory-Stockmayer theories and the delineation of the two-phase regime using our analysis of the sharp changes in ρ. This phase diagram is plotted in the two-parameter space of the concentration of domains along the abscissa and increasing intrinsic affinities along the ordinate. As the intrinsic affinities increase, the widths of the two-phase regimes also increase, and phase separation is realized at lower concentrations of the interacting domains. Figure 4.8b shows comparisons between six separate phase diagrams, each corresponding to a different magnitude of v_{ex}. For a given value of the intrinsic affinity, the width of the two-phase regime increase as the magnitude of the excluded volume decreases. In contrast, the two-phase regime becomes negligibly small as the magnitude of the linker excluded volume increases. In fact, for high linker excluded volumes, the presence of a two-phase regime is only discernible for large intrinsic affinities.

For a specific choice of intrinsic affinities, we shall denote the coexisting concentrations corresponding to the two-phase regime are denoted as c_{cl} and c_{ch}, and the gel point predicted by the Flory-Stockmayer theories as $c_{g,FS}$. It follows that $c_{cl} < c_{g,FS} < c_{ch}$ such that phase separation

will support gelation within the droplet. The width of the two-phase regime is quantified by the gap parameter $g_w = |c_{cl} - c_{ch}|$. The extent of coupling between phase separation and gelation will be governed by the gap parameter $g_c = |c_{cl} - c_g|$, whereas the stability of the gel vis-à-vis the dense liquid will be governed by the gap parameter $g_s = |c_g - c_{ch}|$. Being able to calculate full phase diagrams as shown in figure 4.8 is very helpful because it allows one to make predictions regarding the comparative phase behaviors encoded by changes to linker sequences and / or intrinsic affinities.

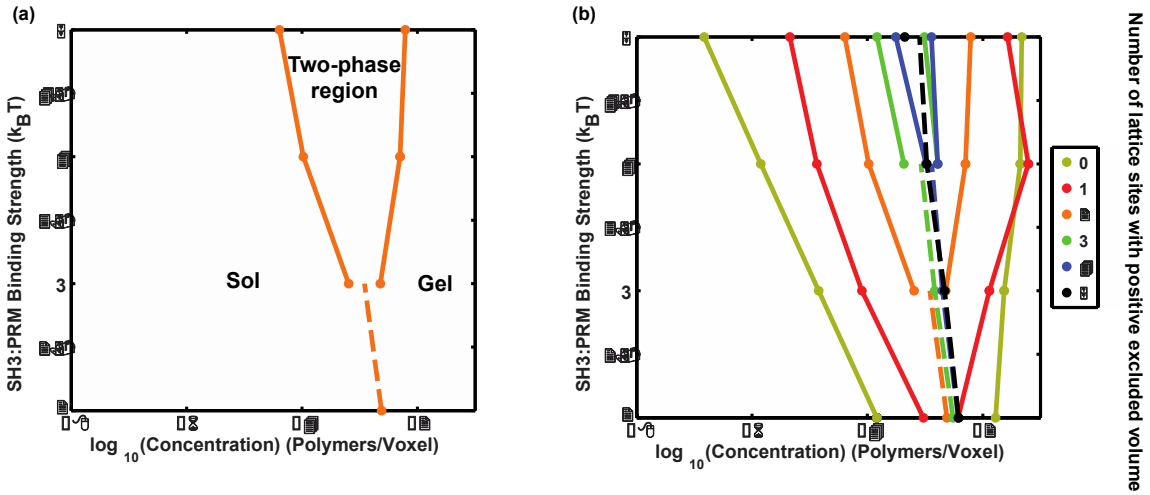


Figure 4.8: Full phase diagrams for multivalent proteins undergoing coupled phase separation and gelation transitions. (a) Two-parameter phase diagram calculated for a 5:5 system with a hybrid five-site linker. For low binding affinities between SH3 domains and PRMs, the system undergoes a sol-gel transition, in accord with Flory-Stockmayer theory when the bulk concentration crosses the gel point, which is delineated by the dashed line. For stronger interactions between SH3 domains and PRMs, the sol-gel transition is preceded by phase separation. This is characterized by a coexistence curve with two arms. Panel (b) shows that the gap parameters g_w, g_c and g_s (see main text) can be modulated by the excluded volume of the linkers connecting SH3 domains and PRMs within multivalent proteins.

4.10 Excluded Volume in Naturally Occurring Linkers is Encoded by Sequence

The preceding analysis raises the question of whether or not the magnitude of the excluded volume can be encoded by sequence features of disordered linkers or through posttranslational modifications? A series of recent studies have shown that the overall sizes, shapes, and amplitudes of conformational fluctuations of disordered proteins are governed by specific sequence-encoded parameters. These include the overall charge contents, the net charge per residues, and the patterning of oppositely charged residues within the linear sequence. The magnitude of the sequence-encoded excluded volume can be quantified through knowledge of the distributions of inter-residue distances within a sequence. This information is accessible through atomistic simulations, single molecule measurements, small angle scattering experiments, or some combination of these methods. The information required to estimate the magnitude of the sequence-encoded excluded volume is embedded in a so-called internal scaling profile for the linker of interest. This profile quantifies the average spatial separation between pairs of residues i, j as a function of the linear sequence separation |j–i|. The corresponding profile for an FRC linker can be calculated either analytically or using a numerical rotational isomeric state approximation. Figure 4.9 shows the internal scaling profile for an FRC linker as a dotted curve. This figure also shows internal scaling profiles that were calculated from all atom simulations of fourteen different disordered linkers that were chosen at random from the human proteome. Similar sequences are often found as linkers between protein-protein and protein-nucleic acid domains. For linkers with positive excluded volume, the value along the ordinate will be greater than the FRC limit along most of the internal scaling profile. As the fraction of

charged residues decreases, the internal scaling profiles converge upon the FRC limit. For sequences that are deficient in charges, the chains compact on themselves, and the sequence-encoded excluded volume is negative. These results emphasize the sequence-encoded tunability of the excluded volume and points to a natural way to tune the coupling between phase separation and gelation.

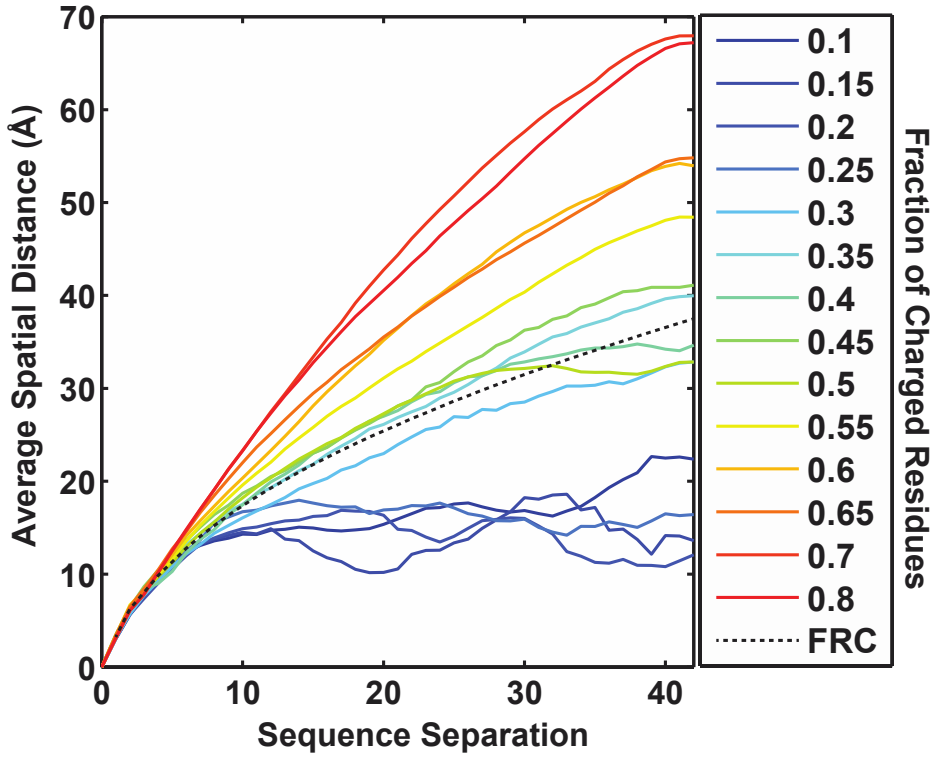


Figure 4.9: The excluded volume parameter in naturally occurring disordered linkers is encoded by sequence. The internal scaling profiles quantify the deviation of sequence-encoded excluded volume from the reference FRC limit – dotted curve. The plot shows the average spatial separation between pairs of residues i, j against the linear sequence separation |j–i|. Results are shown for fourteen naturally occurring linkers of identical length that are extracted for the database of *bona fide* disordered regions. These sequences are often found as linkers between protein-protein and protein-nucleic acid domains. For linkers with positive excluded volume, the value along the ordinate will be greater than the FRC limit along most of the internal scaling profile. As the fraction of charged residues decreases, the internal scaling profiles converge upon the FRC limit. For sequences that are deficient in charges, the chains compact on themselves, and the sequence-encoded excluded volume is negative.

4.11 Discussion

Studies have begun to extract the parameters that are important for proteins that act as scaffolds for the condensation of membraneless organelles. These studies have focused on what types of domains interact with other domains and how this results in recruiting different components into a functional liquid-like body in cells. These proteins are highly enriched in disordered regions and linker regions, regions that are disordered but aren't driving interactions, have been ignored as regions that can strongly influence the phase behavior of the scaffolds. In this work, we have present results from coarse-grained simulations that illustrate how the linkers work synergistically with the interactive domains to control the phase diagram of the system. Our study focuses on two principle first order parameters for a disordered polymer linker: length and excluded volume.

The polymer length is modeled by changing the maximum distance connected domains can separate in our simulations, which has a one-to-one conversion factor to the number of residues in a linker region. Our phase diagrams predict that there is an optimal linker length between domains. For the interaction strength that we focused on, this ideal length was around 20 to 30 residues long. Interestingly, proteins with these ideal linker lengths could drive phase separation up to an order of magnitude better than the classic Flory-Stockmayer theory would suggest.

The polymer excluded volume is modeled by the number of beads that make up the linker. Less intuitively than linker length, the excluded volume has a correlation with the number of charged and proline residues. These residues would much rather interact with the solvent than other polymers which favors strongly avoiding itself. Our phase diagrams predict that the

excluded volume dramatically changes the preference for phase separation to the point where this becomes a useful parameter for tuning between phase separation and a sol-gel transition.

Our results point to linker parameters that should be incorporated into genome searches for scaffolds associated with membraneless organelles. Interestingly, they point to two mechanism for modulating the phase diagram in situ. To drive phase separation, post-translational modifications have the potential to turn on new interaction sites on the protein. In addition to changing the valence, this can significantly change the linker properties through effectively halving the linker length and decreasing the effective excluded volume. To drive dissolution of phase separation, these modifications could cause dramatic increases to the polymeric excluded volume.

We also considered examining the role persistence length plays on the phase diagrams, however, this isn't a physically relevant parameter within the length scale of our coarse-graining. Disordered proteins' behavior is dominated by local geometry constraints and interactions up to around 5 to 7 residues. Beyond this length scale, the general direction of the polymer has no relevant memory. This is equal to or just shy of the length of each lattice site, suggesting that our simulation does not have the resolution where persistence length is relevant to model. Additionally, there is little evidence that unstructured regions have much diversity in their persistence lengths, nor do we have a clear design principle to guide designing or predicting sequences with large variations in persistence length.

While the model was inspired by proteins with ordered binding domains and disordered linkers, it is important to explicit point out that there isn't anything specific in our model that excludes our results from modeling equally well RNA and single-stranded DNA. Both have persistence lengths on similar length scales as disordered proteins and have both interactive and

linker regions encoded by their sequence and chemical modifications. However, special care should be considered when extrapolating these results to double stranded DNA where the persistence length far exceeds these disordered polymers.

4.12 Methods and Analysis

Design of the lattice model and interaction matrix: The multivalent proteins were modeled using a coarse-grained bead-tether model (Fig. 4.3). All simulations were performed on 3-dimensional cubic lattices with periodic boundary conditions. Molecules on the lattice have multiple interaction domains, which refer to either SH3 domains or PRMs that are tethered together via disordered linkers. The number of interaction domains on a given protein is its valence. Each SH3 domain or PRM occupies a distinct lattice site.

The interaction matrix includes the following terms: Each interaction domain (SH3 domain or PRM) or explicitly modeled linker bead has a finite excluded volume such that each lattice site may have only one domain or linker bead. All other interactions are nearest neighbor interactions such that adjacent sites x and y on the lattice are assigned an interaction energy ε_{xy} in units of k_BT, where k_B is Boltzmann's constant and T is the simulation temperature. We designate lattice sites occupied by SH3 domains using the letter S; sites occupied by PRMs by the letter P; and sites occupying linker beads by the letter L. In the default model, $\varepsilon_{SS} = \varepsilon_{PP} = \varepsilon_{LL} = \varepsilon_{SL} = \varepsilon_{PL} = 0$ and $\varepsilon_{SP} = -2k_BT$.

Design of Monte Carlo moves for simulating the phase behavior of multivalent proteins: Five types of moves were deployed to evolve the system. (i) In addition to occupying adjacent lattice sites, a pair of interacting domains in a bound state if and only if this is specified by the interaction state of the domains. Accordingly, one of the moves randomly changes the interaction

state of a domain without changing lattice positions. (ii) The torsional state of an end module that is tethered on one side is altered and a new interaction state is chosen at random. This attempts to move the module to a new location that is within tethering range of the linker, which is the maximum allowable length for the linker. If the module is an interaction domain, then this move additionally changes the interaction state of the domain similar to move 1. (iii) Crankshaft motions are applied to modules tethered on both sides. The module is moved to a new location that is within tethering range of all linkers that connect to the module in question. This is followed by randomly choosing a new interaction state if the module is an interaction domain. (iv) This move involves the collective translation of all modules that are part of a connected network. The latter is calculated by analyzing the list of all proteins that are connected through interacting domains. An arbitrary translation in any direction is then attempted. (v) Finally, individual chains are allowed to undergo reptation via a slithering motion of a protein by removing an end domain and its linker and appending it to the other end. The domain and linker are placed in a random position that maintains the tether ranges. After the new position has been assigned, the interaction state of the domain is randomly assigned.

Acceptance and rejection of Monte Carlo moves: If a move results in placement of a domain or module on a site that is already occupied, then the move is rejected. For rotational, torsional, crankshaft, and reptation moves, the moves that do not lead to steric overlap with occupied sites are accepted according to the modified Metropolis criterion *viz.*, $\min\left\{1, w\exp\left(-\Delta E\right)\right\}$. Here, ΔE is the change in the energy of the system that results from the proposed move. The energy is normalized with respect to k_BT. The parameter w is set based on the proposed type of move. For rotational moves, w=1; for torsional and crankshaft moves, $w = \left(\frac{N_p}{N_c}\right)$, where N_p and N_c are the

number of possible interacting states in the proposed and current states, respectively; finally, for reptation moves, $w = \left(\frac{N_p V_p}{N_c V_c} \right)$, where N_p and N_c are the number of possible interacting states in the proposed and current states, respectively whereas V_p and V_c are the total number of conformations the domain and linker could be placed in the proposed state and current state respectively. These modifications to the standard Metropolis Monte Carlo acceptance criterion ensure the preservation of microscopic reversibility. The translation of a connected network does not create or destroy interactions, nor does it move the relevant linkers. Therefore, the proposed translational moves are always accepted if the move does not lead to steric overlaps.

Flory-Stockmayer Theory: The percolation threshold for high valence polymers can be estimated by analytical methods, one of which is the Flory-Stockmayer theory. In this theory the important parameters are the effective valence or number of interacting modules of the polymers, V, and the binding fraction, f. For a specific protein that is bound to a hypothetical cluster or network the mean number of *additional* proteins recruited to the network, ε, can be expressed as $\varepsilon = (V - 1)f$.

It uses the fraction bound minus 1 because one module is already committed to binding to the network. In a system with two types of proteins the mean number of additional proteins recruited to the network can be expressed as the product of the two protein's values,

$\varepsilon = \varepsilon_a \varepsilon_b = (V_a - 1)f_a(V_b - 1)f_b$

When ε is greater than 1, on average, each protein that binds with the network brings with it more than one additional protein, expanding the network. These proteins are also on average bringing more than one more protein, expanding the network even more. This cascades into an infinitely large cluster of proteins. However, if ε is less than 1 then the proteins added are more

likely to end the network than propagate it so the network is destined to eventually terminate.

For our system, we can calculate the fraction of interactions through the K_D,

$$K_D = \frac{(a-[ab])(b-[ab])}{[ab]}$$

$$[ab] = \frac{(a+b+K_D-\sqrt{(a+b+K_D)^2-4ab})}{2}$$

$$f_a = \frac{[ab]}{a} = \frac{(a+b+K_D-\sqrt{(a+b+K_D)^2-4ab})}{2a}$$

$$\varepsilon = \frac{\left(a+b+K_D-\sqrt{(a+b+K_D)^2-4ab}\right)^2}{4ab}(V_a-1)(V_b-1)$$

We can solve for the critical concentration of module a as a function of b by setting ε=1.

$$a = \frac{b+\gamma^2 b-2\gamma K_D \pm(\gamma+1)\sqrt{b^2(\gamma-1)^2-4\gamma K_D}}{2\gamma}$$

where

$$\gamma \stackrel{\text{def}}{=} (V_a-1)(V_b-1).$$

Alternatively, we can calculate the critical concentration for the equimolar case where a=b,

$$a = \frac{K_D\sqrt{\gamma}}{(1-\sqrt{\gamma})^2}$$

Production runs to generate phase diagrams: For a majority of the simulations, except those where finite size artifacts were queried or the binding affinities were titrated, the interaction energy between adjacent sites with SH3 domains and PRMs was set to $-2k_BT$. In every system, there were 2.4×10^3 interaction domains. Concentrations of domains were titrated by changing the number of lattice sites. Each simulation was run for 5×10^9 steps and the average over the last half was used to calculate the size of the largest connected network.

In order to query the onset of a gelation transition, we used the largest connected network in a simulation, which is the same as the fraction of molecules that make up the largest connected

cluster within the system. We designate this as ϕ_c. The value of ϕ_c that is associated with crossing the critical concentration for percolation, defined as the gel point, is determined by comparing the largest connected network from a randomly generated network to the critical concentration predicted by Flory-Stockmayer theory. Here, the number of nodes in the random network is set to the number of interaction domains used in the lattice simulations. The random network was generated for stoichiometric concentrations of complementary domains. For each domain of type A, a random number was compared to the gross probability that an individual domain would be interacting with a domain of type B, *f*. If the random number was less than *f*, a partner was chosen randomly among the domains of type B that do not already have a binding partner. The results shown in figure 4.5 are obtained by averaging over 10^3 replicates.

Calculation of Phase Boundaries: We utilized ρ as the order parameter for differentiating between the sol:gel transition and phase separation. At the critical point we calculated the concentrations of the polymer rich and polymer poor phases for the two arms in the phase transition regime by using the simplifying assumption that the polymer rich phase is a uniform density sphere and the polymer poor phase has a uniform density in the lattice except, obviously, in the sphere occupied by the polymer rich phase. We can solve for the radius of the phase separated sphere by solving for the physically relevant root of

$$\frac{12}{25}\pi N_T r_N^5 - \frac{4}{3}N_T R_g^2 r_N^3 - \frac{9}{25}N_N L^3 r_N^2 + \frac{(N_N - N_T)L^5}{4} + N_T L^3 R_g^2 = 0,$$

where N_T is the number of proteins in the simulation, N_N is the number of proteins in the largest network, Rg is the radius of gyration of all the proteins in the simulation, L is the lattice length on a side, and r_N is the radius of the polymer dense phase.

This equation was never observed to had more than one real root that could fit inside the lattice for any of our simulations. The phase boundaries were then calculated using

$$c_{cl} = \frac{\left(N_T - N_N\right)}{\left(L^3 - \frac{4}{3}\pi r_N^3\right)}, \text{ and } c_{ch} = \frac{3N_n}{4\pi r_N^3}.$$

Quantification of Finite Sampling: In addition to starting simulations in the random coil state, we also quantified the phase diagrams when the simulations started in a dense phase separated state. For each simulation we equilibrated the proteins in the gel state in a box size of 34 lattice for 5×10^9 steps. The resulting conformation was then used as the simulation's initial conditions in a larger box by expanding the lattice boundary to achieve the desired concentration. For proteins that span the periodic boundary, the first domain was used as the protein's reference for picking which protein image to keep. These initial conditions reproduced the critical concentrations as a function of valence and length (Fig. 4.10).

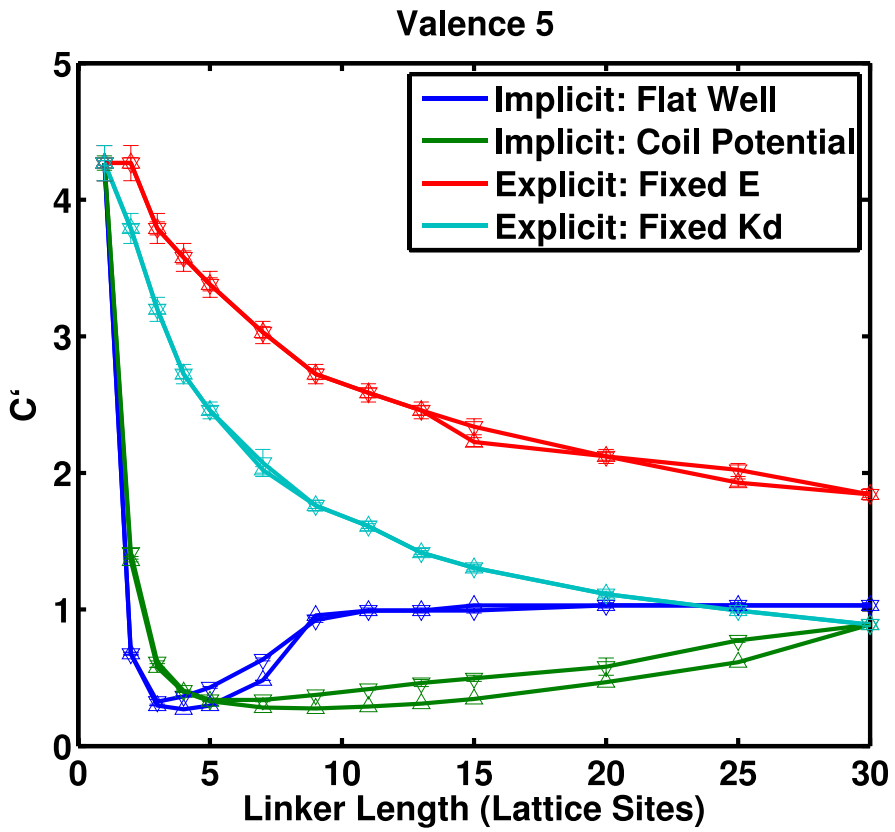


Figure 4.10: Analysis of Errors from Finite Sampling. Linker length of critical concentration from coil (upward facing triangle) and phase separated (downward facing triangle) initial conditions. The error bars are the min and max from two independent calculations of the critical concentration while the triangle is the value from averaging all the results from both simulations. For most simulations, the two initial conditions gave identical results. For all simulations that gave different results by more than a single lattice unit, the phase separated initial conditions gave a higher critical concentration as one would expect.

4.13 Acknowledgments

We are grateful to Alex Holehouse, Anu Mittal, Kiersten Ruff, and Jason Wagoner for stimulating discussions. Grants from the National Science Foundation (MCB1614766 to RVP), the National Institutes of Health (RO1-GM56322 to MKR) and the Howard Hughes Medical Institute (MKR) supported this work. TSH is a graduate student scholar of the Center for Biological Systems Engineering at Washington University in St. Louis.

4.14 References

1. Banani, S.F., et al., *Biomolecular condensates: organizers of cellular biochemistry.* Nat Rev Mol Cell Biol, 2017.
2. Brangwynne, C.P., et al., *Germline P granules are liquid droplets that localize by controlled dissolution/condensation.* Science, 2009. **324**(5935): p. 1729-32.
3. Hyman, A.A. and C.P. Brangwynne, *Beyond stereospecificity: liquids and mesoscale organization of cytoplasm.* Dev Cell, 2011. **21**(1): p. 14-6.
4. Lee, C.F., et al., *Spatial organization of the cell cytoplasm by position-dependent phase separation.* Phys Rev Lett, 2013. **111**(8): p. 088101.
5. Hyman, A.A., C.A. Weber, and F. Julicher, *Liquid-liquid phase separation in biology.* Annu Rev Cell Dev Biol, 2014. **30**: p. 39-58.
6. Li, P., et al., *Phase transitions in the assembly of multivalent signalling proteins.* Nature, 2012. **483**(7389): p. 336-40.
7. Su, X., et al., *Phase separation of signaling molecules promotes T cell receptor signal transduction.* Science, 2016. **352**(6285): p. 595-9.
8. Feric, M., et al., *Coexisting Liquid Phases Underlie Nucleolar Subcompartments.* Cell, 2016. **165**(7): p. 1686-97.
9. Zhu, L. and C.P. Brangwynne, *Nuclear bodies: the emerging biophysics of nucleoplasmic phases.* Curr Opin Cell Biol, 2015. **34**: p. 23-30.
10. Mitrea, D.M., et al., *Nucleophosmin integrates within the nucleolus via multi-modal interactions with proteins displaying R-rich linear motifs and rRNA.* Elife, 2016. **5**.
11. Parry, B.R., et al., *The bacterial cytoplasm has glass-like properties and is fluidized by metabolic activity.* Cell, 2014. **156**(1-2): p. 183-94.

12. Munder, M.C., et al., *A pH-driven transition of the cytoplasm from a fluid- to a solid-like state promotes entry into dormancy.* Elife, 2016. **5**.

13. Ramaswami, M., J.P. Taylor, and R. Parker, *Altered ribostasis: RNA-protein granules in degenerative disorders.* Cell, 2013. **154**(4): p. 727-36.

14. Saha, S., et al., *Polar Positioning of Phase-Separated Liquid Compartments in Cells Regulated by an mRNA Competition Mechanism.* Cell, 2016. **166**(6): p. 1572-1584.e16.

15. Nott, T.J., et al., *Phase transition of a disordered nuage protein generates environmentally responsive membraneless organelles.* Mol Cell, 2015. **57**(5): p. 936-47.

16. Lee, C., P. Occhipinti, and A.S. Gladfelter, *PolyQ-dependent RNA-protein assemblies control symmetry breaking.* J Cell Biol, 2015. **208**(5): p. 533-44.

17. Jain, S. and R. Parker, *The discovery and analysis of P Bodies.* Adv Exp Med Biol, 2013. **768**: p. 23-43.

18. Banani, S.F., et al., *Compositional Control of Phase-Separated Cellular Bodies.* Cell, 2016. **166**(3): p. 651-63.

19. Wheeler, J.R., et al., *Distinct stages in stress granule assembly and disassembly.* Elife, 2016. **5**.

20. Brangwynne, C.P., P. Tompa, and R.V. Pappu, *Polymer physics of intracellular phase transitions.* Nat Phys, 2015. **11**(11): p. 899-904.

21. Papayannopoulos, V., et al., *A polybasic motif allows N-WASP to act as a sensor of PIP(2) density.* Mol Cell, 2005. **17**(2): p. 181-91.

22. Sallee, N.A., et al., *The pathogen protein EspF(U) hijacks actin polymerization using mimicry and multivalency.* Nature, 2008. **454**(7207): p. 1005-8.

23. Tsygankov, A.Y., et al., *Beyond the RING: CBL proteins as multivalent adapters.* Oncogene, 2001. **20**(44): p. 6382-402.

24. Bunnell, S.C., et al., *Persistence of cooperatively stabilized signaling clusters drives T-cell activation.* Mol Cell Biol, 2006. **26**(19): p. 7155-66.

25. Sriram, S.M., et al., *Multivalency-assisted control of intracellular signaling pathways: application for ubiquitin- dependent N-end rule pathway.* Chem Biol, 2009. **16**(2): p. 121-31.

26. Dam, T.K. and C.F. Brewer, *Multivalent lectin-carbohydrate interactions energetics and mechanisms of binding.* Adv Carbohydr Chem Biochem, 2010. **63**: p. 139-64.

27. Liu, F. and K.J. Walters, *Multitasking with ubiquitin through multivalent interactions.* Trends Biochem Sci, 2010. **35**(6): p. 352-60.

28. Preissner, K.T. and U. Reuning, *Vitronectin in vascular context: facets of a multitalented matricellular protein.* Semin Thromb Hemost, 2011. **37**(4): p. 408-24.

29. Weise, K., et al., *Membrane-mediated induction and sorting of K-Ras microdomain signaling platforms.* J Am Chem Soc, 2011. **133**(4): p. 880-7.

30. Wilson, B.S., J.M. Oliver, and D.S. Lidke, *Spatio-temporal signaling in mast cells.* Adv Exp Med Biol, 2011. **716**: p. 91-106.

31. Husnjak, K. and I. Dikic, *Ubiquitin-binding proteins: decoders of ubiquitin-mediated cellular functions.* Annu Rev Biochem, 2012. **81**: p. 291-322.

32. Bai, J.J., et al., *Ataxin-3 is a multivalent ligand for the parkin Ubl domain.* Biochemistry, 2013. **52**(42): p. 7369-76.

33. Falkenberg, C.V., M.L. Blinov, and L.M. Loew, *Pleomorphic ensembles: formation of large clusters composed of weakly interacting multivalent molecules.* Biophys J, 2013. **105**(11): p. 2451-60.

34. Nair, S.S., D.Q. Li, and R. Kumar, *A core chromatin remodeling factor instructs global chromatin signaling through multivalent reading of nucleosome codes.* Mol Cell, 2013. **49**(4): p. 704-18.

35. Nussinov, R. and H. Jang, *Dynamic multiprotein assemblies shape the spatial structure of cell signaling.* Prog Biophys Mol Biol, 2014. **116**(2-3): p. 158-64.

36. Satav, T., J. Huskens, and P. Jonkheijm, *Effects of Variations in Ligand Density on Cell Signaling.* Small, 2015. **11**(39): p. 5184-99.

37. Flory, P.J., *Molecular Size Distribution in Three Dimensional Polymers. I. Gelation1.* Journal of the American Chemical Society, 1941. **63**(11): p. 3083-3090.

38. Flory, P.J., *Constitution of Three-dimensional Polymers and the Theory of Gelation.* The Journal of Physical Chemistry, 1942. **46**(1): p. 132-140.

39. Stockmayer, W.H., *Theory of Molecular Size Distribution and Gel Formation in Branched‐Chain Polymers.* The Journal of Chemical Physics, 1943. **11**(2): p. 45-55.

40. Banjade, S. and M.K. Rosen, *Phase transitions of multivalent proteins can promote clustering of membrane receptors.* Elife, 2014. **3**.

41. Banjade, S., et al., *Conserved interdomain linker promotes phase separation of the multivalent adaptor protein Nck.* Proc Natl Acad Sci U S A, 2015. **112**(47): p. E6426-35.

42. Flory, P.J., *Introductory lecture.* Faraday Discussions of the Chemical Society, 1974. **57**(0): p. 7-18.

43. Flory, P.J., *Thermodynamics of High Polymer Solutions.* The Journal of Chemical Physics, 1942. **10**(1): p. 51-61.

44. Huggins, M.L., *Some Properties of Solutions of Long-chain Compounds.* The Journal of Physical Chemistry, 1942. **46**(1): p. 151-158.

45. Semenov, A.N. and M. Rubinstein, *Thermoreversible Gelation in Solutions of Associative Polymers. 1. Statics.* Macromolecules, 1998. **31**(4): p. 1373-1385.

46. Rubinstein, M. and A.N. Semenov, *Thermoreversible Gelation in Solutions of Associating Polymers. 2. Linear Dynamics.* Macromolecules, 1998. **31**(4): p. 1386-1397.

47. Rubinstein, M. and R.H. Colby, *Polymer Physics*. 2003, Oxford and New York: Oxford University Press.

48. Das, R.K., K.M. Ruff, and R.V. Pappu, *Relating sequence encoded information to form and function of intrinsically disordered proteins.* Curr Opin Struct Biol, 2015. **32**: p. 102-12.

49. Mittal, A., et al., *Hamiltonian Switch Metropolis Monte Carlo Simulations for Improved Conformational Sampling of Intrinsically Disordered Regions Tethered to Ordered Domains of Proteins.* J Chem Theory Comput, 2014. **10**(8): p. 3550-3562.

Chapter 5

Coexisting Liquid Phases Underlie Nucleolar Sub-Compartments

This chapter is adapted from an article[1] published in Cell. Marina Feric, Nilesh Vaidya, Diana M. Mitrea, Lian Zhu, Tiffany M. Richardson, Richard W. Kriwacki, and Clifford P. Brangwynne designed and conducted the experiments. Tyler S. Harmon and Rohit V. Pappu developed the coarse-grained framework. Tyler S. Harmon performed and analyzed the simulations.

5.1 Introduction

The cellular interior is organized into organelles whose structures have evolved to facilitate their functions. The most well known examples are the canonical membrane-bound organelles such as secretory vesicles, the Golgi apparatus, and the endoplasmic reticulum. However, many intracellular compartments are membrane-less bodies comprised of RNA and protein, often referred to as RNP bodies; these include stress granules and processing bodies in the cytoplasm, and Cajal bodies and nucleoli in the nucleus, among many others. Despite their lack of a delimiting membrane, these organelles nevertheless maintain a coherent size and shape, with a well-defined boundary that compartmentalizes different types of proteins and RNA. By concentrating molecules within a micro-compartment, while allowing dynamic molecular interactions, these RNP bodies may function to control reaction efficiencies much like conventional membrane bound cytoplasmic organelles [2-4].

Many of these RNP bodies exhibit liquid-like biophysical properties, and growing evidence suggests they assemble via liquid-liquid phase separation [5-8]. Intracellular phase

transitions can result in switch-like changes in molecular organization and the spontaneous formation of micron-scale membrane-less organelles. Such behavior is reminiscent of well-known *in vitro* observations in protein crystallization, where soluble proteins are observed to condense into concentrated liquid phases or crystalline solid phases. A number of recent papers suggest that intrinsically disordered proteins or low complexity sequences (IDP/LCS) drive phase transitions underlying assembly of the nucleolus [9], stress granules [10-12], P granules and nuage bodies [13, 14], and nuclear speckles [15].

The liquid-like nature of the nucleolus may facilitate its function in ribosome biogenesis. The nucleolus forms around regions of chromosomes containing stretches of tandem ribosomal DNA (rDNA) gene repeats, known as nucleolar organizer regions (NORs). In most eukaryotes (including human, *X. laevis*, and *C. elegans*) a precursor ribosomal RNA (rRNA) transcript is generated from the rDNA gene, and contains each of the co-transcribed 18S, 5.8S, and 28S rRNA subunits, separated by two intervening transcribed sequences (Fig. 5.1A). The nucleolus may facilitate increased reaction rates by locally concentrating rRNA processing factors involved in transforming the precursor rRNA transcript into individual rRNA subunits. Due to its role in producing this protein translational machinery, the structure and function of the nucleolus are intimately connected with cell growth and size homeostasis [16, 17].

Despite the biological importance of the nucleolus, a mechanistic biophysical understanding of its assembly and internal organization is lacking. The simplest picture of the nucleolus as a unitary liquid phase body becomes difficult to reconcile with its well-known complex and multi-component nature. Indeed, in addition to the various types of RNA in the nucleolus, the nucleolar proteome consists of hundreds of different proteins that are segregated into at least three distinct compartments [18]. This layered tripartite organization consists of the

fibrillar center (FC), where the RNA polymerase I (POL1) machinery is active; the dense fibrillar component (DFC) that is enriched in the protein fibrillarin (FIB1); and the granular component (GC) that is enriched in the protein nucleophosmin (NPM1/B23) (Fig. 5.1A,B). This multi-layered structure is not unique to the nucleolus, as stress granules and other liquid-like RNP bodies exhibit similar "core-shell" structuring [19, 20].

The multi-layered structure of the nucleolus is thought to facilitate assembly-line processing of rRNA. Nascent rRNA transcripts undergo sequential processing steps by enzymes that localize to the distinct compartments, ultimately exiting the nucleolus, and being exported for final ribosome assembly in the cytoplasm (Fig. 5.1A). Although recent work has shown that the entire nucleolus can exhibit active liquid-like properties [21] and its assembly may represent a type of liquid-liquid phase transition [8], reconciliation of these findings with the multi-layered structure of the nucleolus has proven elusive. Indeed, if the nucleolus is a type of liquid, what mechanism prevents the three components from mixing and fusing to form a single liquid phase?

Here, we uncover a physical mechanism for intranucleolar organization: differences in miscibility between proteins from different nucleolar compartments keep the compartments phase separated, giving rise to the layered, multiphase droplet nature of nucleoli. By isolating protein domains from key nucleolar proteins, we provide evidence for a molecular mechanism whereby intrinsically disordered regions (IDRs) help drive protein condensation into droplets, while associated RNA binding domains confer sub-compartment specificity by making the two droplet phases immiscible with one another.

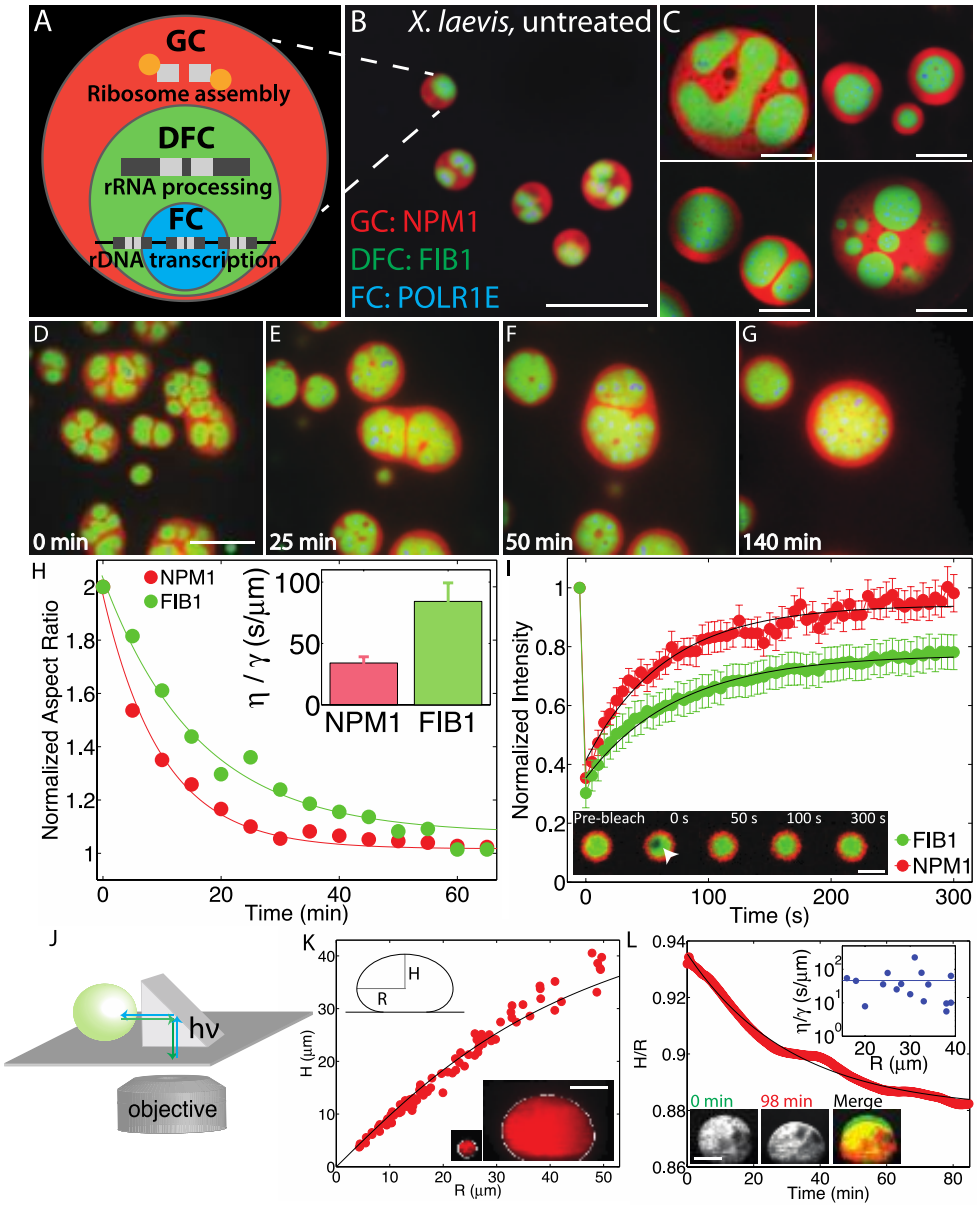


Figure 5.1: Liquid-like behavior of biophysically distinct nucleolar sub-compartments. (A) Schematic diagram of ribosome biogenesis in nucleolus. (B) Nucleoli in an untreated *X. laevis* nucleus. Scale bar = 20 μm. For all images, granular component (GC) is visualized with NPM1 (red), dense fibrillar component (DFC) with FIB1 (green), and fibrillar center (FC) with POLR1E (blue). (C) Examples of nucleoli after coarsening in *X. laevis* nuclei treated with Lat-A. Scale bar = 20 μm. (D-G) Time-course of nucleolar component fusion after actin disruption by Lat-A. Scale bar = 20 μm. (H) Normalized aspect ratio vs. time for nucleolar components fusing after actin disruption. Inset shows η/γ for 59 nucleoli analyzed from 20 nuclei. (I) FRAP recovery curves for NPM1 (red) and FIB1 (green) in *X. laevis* oocytes. Inset: FRAP of FIB1-labeled DFC (green). Scale bar = 5 μm. (J) Schematic illustrating XZ imaging with a right angle prism. (K) Height, H, vs. radius, R, of different sized nucleoli at steady-state (91 nucleoli, from 61 nuclei). Black line is the fit from the average surface tension for all nucleoli. Bottom inset: example of the shape of a small vs. large nucleolus. Scale bar = 40 μm. (L) Example of nucleolar height to radius ratio, H/R, vs. time for one nucleolus deforming under gravity. Black line is an exponential fit. Top inset: η/γ for 16 nucleoli in 16 different nuclei. Blue line indicates average. Bottom inset shows example deforming nucleolus: Scale bar = 40 μm.

5.2 Nucleolar Sub-Compartments Behave as Liquid-Like Phases in Vivo

To gain insight into the biophysical assembly principles underlying nucleolar structure, we took advantage of the numerous large nucleoli, ranging in size from 1-10 microns, found in the nucleus (germinal vesicle, GV) of *X. laevis* oocytes (Fig. 5.1B). This system is also convenient because the nucleoli are extra-chromosomal, forming around amplified stretches of rDNA, allowing us to disentangle the confounding effects of somatic chromosome architecture on nucleolar structure. Nucleoli in *X. laevis* oocytes will fuse with one another when brought into contact [21], but the frequency of such coalescence events is slowed significantly by the presence of a nuclear actin network (Fig. 5.2A-D) [22, 23].

We visualized intranucleolar organization by labeling individual components of nucleoli with fluorescent fusion proteins as follows: granular component with nucleophosmin (GC, NPM1::Cerulean), dense fibrillar component with fibrillarin (DFC, FIB1::RFP), and the fibrillar center with RNA polymerase 1E (FC, GFP::POLR1E). To test the role of nuclear actin in organizing these nucleolar substructures, we utilized the actin-disrupting compound Latrunculin-A (Lat-A). As shown previously, the entire nucleolus undergoes liquid-like coalescence events, which can be seen by fusion of the NPM1 (GC) region of one nucleolus with the NPM1 (GC) region from another nucleolus (Fig. 5.1D-G). The FC regions (POLR1E) rarely came into close contact with each other, and consequently, we did not observe fusion between multiple FCs, but we did see rearrangements into more spherical FC domains. Strikingly, however, we typically observe that the FIB1 (DFC) cores from one nucleolus will fuse when in close proximity with FIB1 (DFC) cores from a different nucleolus (Fig. 5.1D-G, 5.2E-L). These DFC regions, which

were initially irregular in shape, would round up and coalesce, exhibiting classic liquid-like behavior. After long times (~1 hour), these coalescence events ultimately resulted in DFCs located in the center of the nucleolus, surrounded by one continuous phase of GC, representing an outer-most enveloping compartment (Fig. 5.1C).

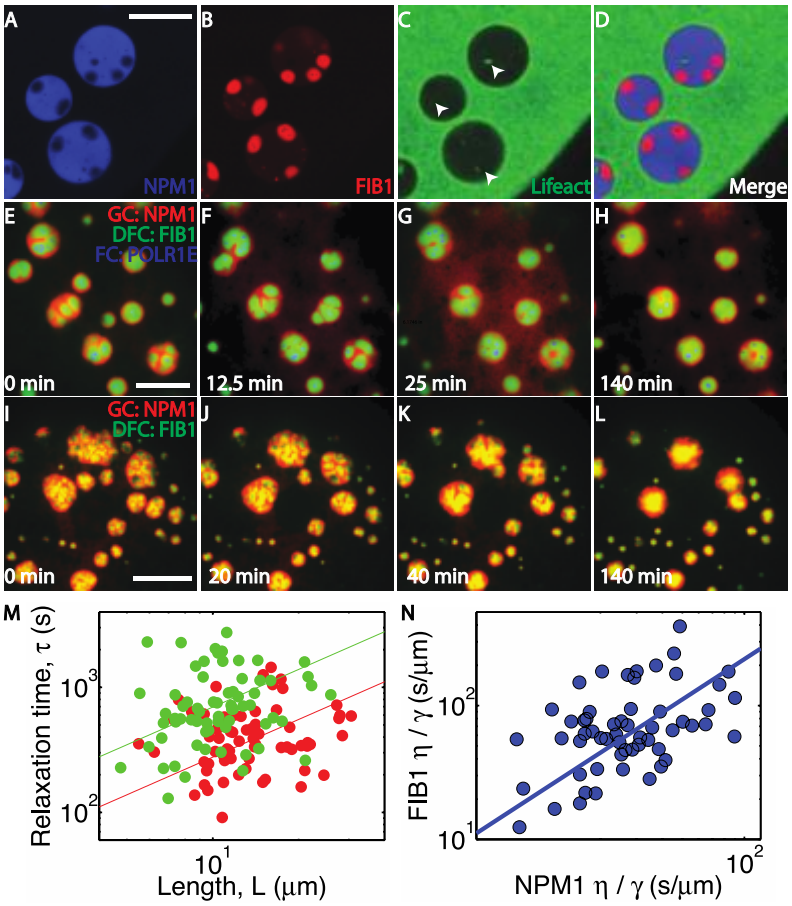


Figure 5.2: Liquid-like behavior of biophysically distinct nucleolar sub-compartments. (A-D) Nucleoli embedded within a nuclear actin network in *X. laevis* nuclei. Granular component is visualized in blue with NPM1::Cerulean (A), dense fibrillar component is visualized with FIB1::RFP (B), and nuclear actin is visualized with Lifeact::GFP (C). Final panel is a merge of the previous three panels (D). Scale bar = 20 μm. (E-H) Nucleoli coarsening after actin disruption visualizing the granular component (NPM1, red), dense fibrillar component (FIB1, green), and the fibrillar centers (POLR1E, blue). Scale bar = 20 μm. (I-L) Nucleoli coarsening after actin disruption visualizing the granular component (NPM1, red) and dense fibrillar component (FIB1, green). Scale bar = 50 μm. (M) Relaxation time, τ, versus nucleolar compartment size, L, for NPM1 (red) and FIB1 (green). (N) Inverse capillary velocity (η/γ) of FIB1 versus NPM1, to show correlation between nucleoli sampled from different nuclei.

5.3 In Vivo Sub-Compartments have Different Biophysical Properties

To gain insight into the biophysical properties of different nucleolar sub-compartments, we quantitatively analyzed fusion events. We found that homotypic fusion between the GC (NPM1) or DFC (FIB1) occurs by exponential relaxation to a single larger spherical shape; this is characteristic of coalescing liquid droplets and can be used to determine the ratio of droplet viscosity, η, to surface tension, γ, known as the inverse capillary velocity: η/γ [21] (Fig. 5.1H). We find that FIB1-labeled DFC tends to exhibit slower fusion dynamics, with a larger value η/γ = 80 ± 15 s/μm (mean ± s.e.m) compared to NPM1-labeled GC with $\eta/\gamma = 30 \pm 5$ s/μm (mean ± s.e.m.) (Fig. 5.1H inset, 5.2M,N). This behavior suggests that these nucleolar components behave as distinct liquid-like phases within the nucleolus, with different properties that could underlie nucleolar structural organization.

Interfaces represent sharp concentration gradients, and surface tension, with units of free energy per unit area, is the energetic cost of increasing the interfacial area. Surface tension is a key parameter that governs how two different droplets interact with one another. We therefore hypothesized that different surface tension values of the nucleolar sub-phases could explain their immiscibility and multi-layered organization. To measure the surface tension of the outermost GC compartment (i.e. interfacial energy of GC/nucleoplasm interface), we disrupted actin and allowed nucleoli to fuse for several hours. This resulted in a single large coalesced nucleolar droplet, which becomes measurably flattened at the bottom of the nucleus due to gravity. Since this flattening is resisted by surface tension, we could use a right-angle prism to measure the (XZ) shape of the droplet and determine the surface tension: $4 \pm 1\times10^{-7}$ N/m (mean ± s.e.m)

(Fig. 5.1J-K); this value is very low, roughly five orders of magnitude lower than water-oil surface tension values [24], but is comparable to values reported for colloidal liquids [25]. Complementary measurements of the droplet flattening timescale combined with droplet fusion measurements are consistent with this low value for the surface tension of the NPM1-rich GC compartment (Fig. 5.1L).

To confirm the liquid-like dynamics of the nucleolar sub-compartments, we performed fluorescence recovery after photobleaching (FRAP) experiments of *X. laevis* nucleoli *in vivo* (Fig. 5.1I). NPM1 exhibits fast dynamics with a nearly complete recovery on a timescale of $\tau = 64 \pm 8$ s for a bleach spot of 1 μm, consistent with the expected response for diffusion within a liquid. However, FIB1 recovery was slightly slower ($\tau = 75 \pm 7$ s). Moreover, the recovery of FIB1 was not complete, but only reached ~80%; this suggests that the DFC sub-compartment may not be a simple liquid, but instead may exhibit partially solid-like properties (i.e., viscoelasticity).

5.4 Purified FIB1 and NPM1 Can Phase Separate into Droplets Similar to Nucleoli in Vivo

To gain further insight into how nucleolar proteins could give rise to distinct liquid-like nucleolar sub-phases, we purified recombinant FIB1 and NPM1, and studied their behavior *in vitro*. Consistent with our previous work, we find that FIB1::GFP (hereafter referred to simply as FIB1) can phase separate *in vitro* under near physiological protein and salt concentrations [9], giving rise to condensed liquid-phase droplets that are roughly 50-fold more concentrated with protein than the surrounding dilute phase (Fig. 5.3A). Indeed, in the presence of 5 μg/mL rRNA and 150 mM NaCl, FIB1 condenses into droplets at a protein concentration of roughly 600 nM.

FIB1 can phase separate even in the presence of non-specific poly-U50 and heparin, suggesting that electrostatic interactions contribute to droplet assembly (Fig. 5.4A). NPM1 has also recently been demonstrated to undergo phase separation into liquid-like droplets (Fig. 5.3B) [4]. However, at 150 mM NaCl, NPM1 requires significantly higher concentrations of protein (2 μM) and rRNA (100 μg/mL). Moreover, phase separation of NPM1 required rRNA and cannot be induced by the addition of heparin or poly-U50.

Given the distinct biophysical properties of the nucleolar sub-compartments, we hypothesized that the two different types of *in vitro* droplets would also exhibit different material properties. As with the *in vivo* sub-compartments, *in vitro* droplets undergo homotypic fusion events when brought into close contact, but do so with markedly different time scales: FIB1 droplets typically take roughly a hundred times longer than NPM1 droplets of comparable size to coalesce and relax into a single larger sphere (Fig. 5.3C). Using an analysis similar to that performed *in vivo*, we find that FIB1 has an inverse capillary velocity of $\eta/\gamma = 40 \pm 10$ s/μm (95% confidence interval), comparable to that measured *in vivo* (Fig. 5.4C). Also mirroring the *in vivo* data, purified NPM1 droplets have a lower value of 0.30 ± 0.07 s/μm (Fig. 5.3D, 5.4C). A series of experiments confirmed that the presence of the GFP tag does effect droplet properties, but does not qualitatively impact our findings (Fig. 5.4B,C).

NPM1 readily formed large droplets *in vitro*, allowing for direct measurement of the surface tension, using a method similar to our *in vivo* set up with the prism (Fig. 5.1J). We estimated a surface tension of $\gamma_{NPM1} = 8 \pm 2{\times}10^{-7}$ N/m (mean ± s.e.m) (Fig. 5.4D). This value is again surprisingly low and on the same order of magnitude as the value obtained for *X. laevis* nucleoli (Fig. 5.4C).

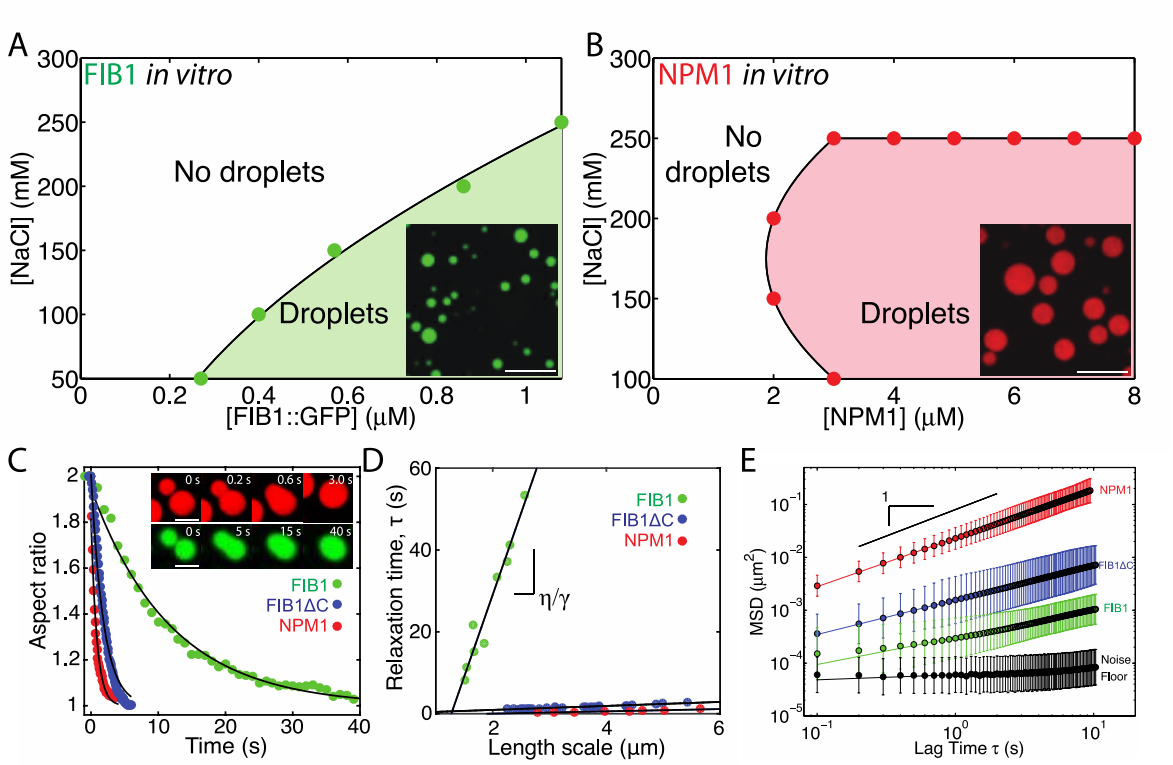


Figure 5.3: Purified nucleolar proteins can phase separate into droplets with different biophysical properties. (A) Phase diagram of purified FIB1 in the presence of 5 μg/ml of rRNA. Inset: FIB1 droplets. Scale bar = 10 μm. (B) Phase diagram of purified NPM1 in the presence of 100 μg/ml of rRNA. Inset: NPM1 droplets. Scale bar = 10 μm. (C) Aspect ratio vs. time for fusing droplets of FIB1 (green), NPM1 (red), and FIB1ΔC (blue). Inset: FIB1 fusing (scale = 2 μm) and NPM1 fusing (scale = 5 μm). (D) Relaxation time versus length scale for fusion data from multiple FIB1 (green), NPM1 (red), and FIB1ΔC droplets (blue). (E) MSD vs. lag time of microrheological probe particles (R=50 nm) embedded in droplets of FIB1 (green), NPM1 (red), or FIB1ΔC (blue); black data points represent the noise floor (black).

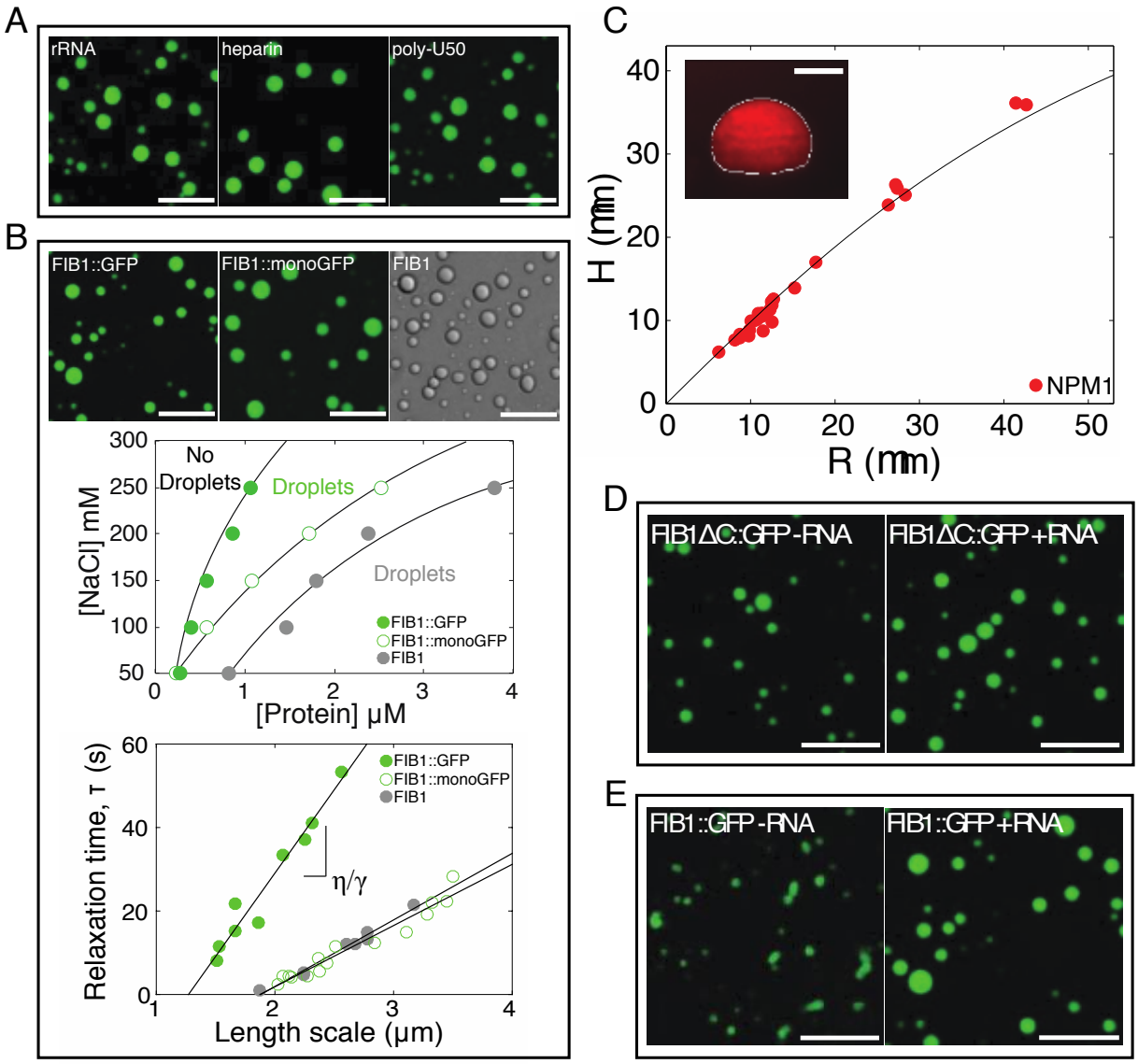


Figure 5.4: Purified nucleolar proteins can phase separate into droplets with different biophysical properties. (A) *In vitro* FIB1 droplets in the presence of 5 µg/ml rRNA (left), or 0.1 mg/ml heparin (middle), or 5 µM poly-U50 (right). Scale bar = 10 µm. (B) Effect of tagging FIB1 on *in vitro* phase separation. The top panel contains *in vitro* FIB1 droplets: FIB1::GFP (left), FIB1::monoGFP (middle), and FIB1 (right). Scale bar = 10 µm. The middle panel is the phase diagram of tagged and untagged *in vitro* FIB1 droplets in the presence of rRNA as a function of protein concentration and salt concentration. The bottom panel is the relaxation time versus length scale for fusion data from multiple FIB1::GFP droplets (closed green), FIB1::monoGFP (GFP with A206K) droplets (open green), and untagged-FIB1 droplets (grey). The slope is the inverse capillary velocity. Inverse capillary velocities of FIB1::GFP is 40 ± 10 s/µm, FIB1::monoGFP is 15 ± 2 s/µm, and FIB1 is 16 ± 2 s/µm. (C) Table summarizing biophysical properties of inverse capillary velocity, viscosity determined from microrheology and surface tension estimated using right angle prism where applicable for NPM1 *in vitro*, NPM1 *in vivo* in *X. laevis*, FIB1::GFP *in vitro*, FIB1::monoGFP *in vitro,* FIB1 (no tag) *in vitro*, FIB1 *in vivo* in *X. laevis*, and FIB1ΔC *in vitro*. (D) Height, *H*, versus radius, *R*, of NPM1 droplets *in vitro* determined from XZ shape profile imaged with a right-angle prism to measure surface tension. Solid black line represents expected trend given average surface tension. Inset shows an XZ view of an NPM1 droplet. Scale bar = 20 µm. (E) *In vitro* FIB1ΔC droplets in the absence (left) or the presence (right) of RNA. (F) *In vitro* FIB1 droplets in the absence (left) or the presence (right) of RNA. FIB1 phase separates into non-liquid, "aggregate-like" structures in the absence of RNA.

5.5 Viscoelasticity and Time-Dependence of in Vitro Droplets

To further shed light on the different properties of the two droplet subtypes, we performed microrheology experiments using the fluctuating motion of probe particles (R = 50 nm) [13, 26]. These data reveal that NPM1 droplets exhibit a diffusive exponent of $\alpha = 0.92 \pm 0.06$, (red symbols, Fig. 5.3E), consistent with that of a simple viscous liquid for which $\alpha = 1$. We can thus calculate a viscosity of NPM1 droplets, $\eta = 0.74 \pm 0.06$ Pa-s, which is several hundred times more viscous than water (Fig. 5.4C). By contrast, probe particle motion in FIB1 droplets is significantly reduced (green symbols, Fig. 5.3E), in agreement with the slowed coalescence dynamics observed with FIB1 droplets. Interestingly, FIB1 droplet microrheology reveals a sub-diffusive exponent ($\alpha = 0.5 \pm 0.1$), which implies that these are not simple viscous liquid droplets, but are instead viscoelastic.

To determine how FIB1 droplet viscoelasticity may arise, we performed FRAP experiments on phase-separated *in vitro* droplets (Fig. 5.5). After 30 minutes of initiating phase separation, we find that NPM1 has near complete recovery ($84 \pm 3\%$) on very short time scales, with a recovery constant of $\tau = 23 \pm 1$ s (Fig. 5.5A, 5.6C); this is consistent with the purely viscous microrheology results, as well as the nearly complete NPM1 FRAP recovery *in vivo*. By contrast FIB1 has low recovery ($37 \pm 2\%$) with a time scale of $\tau = 56 \pm 5$ s (Fig. 5.5B, 5.6C). Such incomplete FRAP recovery is expected for a viscoelastic material, since not all molecules exhibit dynamic, fluid-like exchange. Moreover, this *in vitro* FIB1 behavior agrees well with the incomplete FIB1 FRAP recovery *in vivo*, in both cultured mammalian cells expressing FIB::GFP and amphibian nucleoli (Fig. 5.5F).

To test whether droplet material properties change with time, we performed FRAP experiments on *in vitro* droplets as a function of time. We find that NPM1 always exhibits a near complete recovery, even for droplets that have been sitting for several hours (Fig. 5.5A,D). However, FIB1 FRAP dynamics are strongly impacted by the droplet age. By 2 hours, the percent recovery has dropped by a factor of ~4, to $8 \pm 0.5\%$ (Fig. 5.5B,D); this suggests that these droplets become increasingly solid-like with time, potentially due to the formation of fibers. Consistent with this, we see liquid-like FIB1 droplets evolve into sticky gel-like structures over a 2 hour time period (Fig. 5.6A). Also, we find that replacing the GFP tag with the monomeric GFP (A206K) did not alter the FIB1 aging behavior (Fig. 5.6B). As we describe further below, the N-terminal R/G-rich domain of FIB1, designated as FIB1ΔC, drives phase separation as an autonomous unit. However, it exhibits nearly complete FRAP recovery (Fig. 5.5C, D). Moreover, unlike full length FIB1, the percent recovery is stable over four hours (Fig. 5.5D). Therefore, we conclude that the C-terminal methyltransferase domain of FIB1 plays a key role in promoting viscoelastic maturation of FIB1 droplets *in vitro.*

ATP-dependent active processes have been hypothesized to play an important role in promoting dynamics within cells, in a process known as "active diffusion" [21, 27, 28]. Depleting ATP from *X. laevis* ooctyes or mammalian cells (Fig. 5.5E) had relatively little effect on FRAP recovery of NPM1, with nearly full FRAP recovery, comparable to that seen with *in vitro* NPM1 droplets (Fig. 5.5A, D). By contrast, ATP-depletion resulted in significantly slowed FIB1 dynamics, with longer recovery times in *X. laevis* (Fig. 5.6C), and a 2-3-fold decrease in the percent recovery in both systems (Fig. 5.5F). Moreover, the low percent recovery of FIB1 in ATP-depleted cells (20-40%) is similar to that measured for *in vitro* FIB1 droplets (Fig. 5.5D, F). This suggests that ATP-dependent enzymatic activity is essential for actively maintaining the

fluidity of the aging-prone, FIB1-rich DFC.

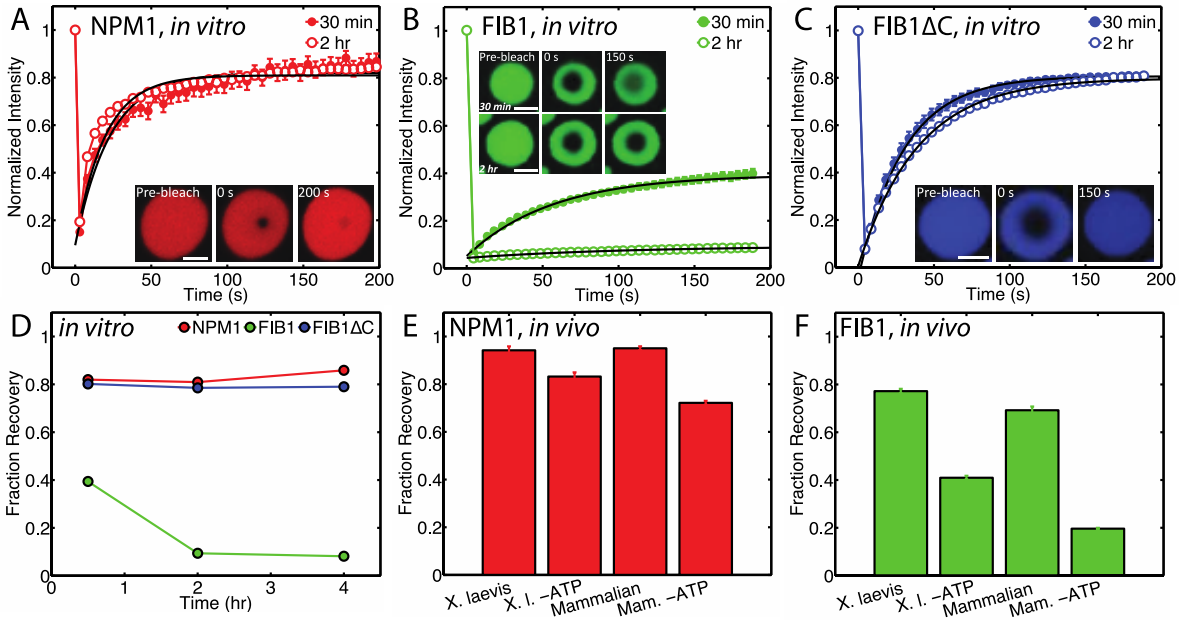


Figure 5.5: Nucleolar protein droplets exhibit liquid-like dynamics, but FIB1 shows evidence for aging. (A-C) FRAP recovery curves for NPM1 (red), FIB1 (green), and FIB1ΔC (blue) droplets, 30 minutes (closed circles) and 2 hours (open squares) after phase separation was initiated. (A) Inset: example FRAP timecourse. Scale bar = 5 μm . (B) Insets: example FRAP timecourses after 30 minutes (top) and 2 hours (bottom). Scale bar = 2 μm. (C) Inset: example FRAP timecourse. Scale bar = 2 μm. (D) Fraction recovery after FRAP experiment as a function of time after phase separation for NPM1 (red), FIB1 (green), and FIB1ΔC (blue) droplets. (E,F) Fraction recovery for NPM1 (E) and FIB1 (F) in *X. laevis* nucleoli and mammalian cell culture nucleoli *in vivo,* for native and ATP depletion conditions.

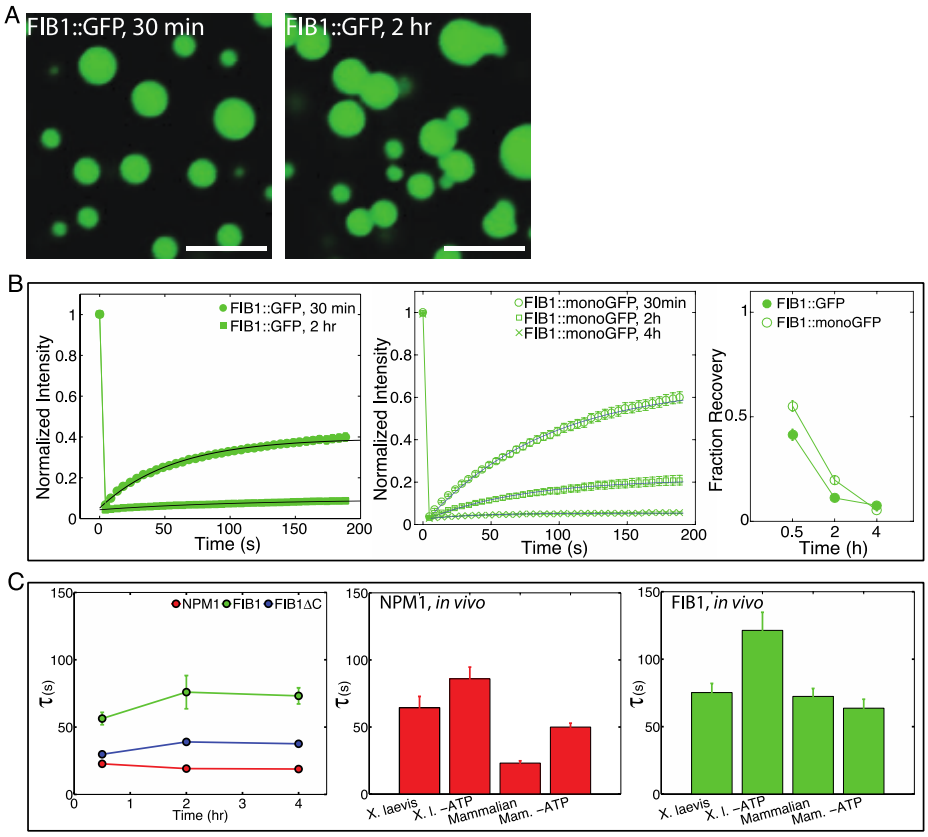


Figure 5.6: Nucleolar protein droplets exhibit liquid-like dynamics, but FIB1 shows evidence for aging. (A) *In vitro* FIB1::GFP droplets at 30 minutes and 2 hours. (B) The left plot is FRAP recovery curves for FIB1::GFP droplets at 30 minutes (closed circles) and 2 hours (closed squares) after phase separation was initiated. The middle plot is FRAP recovery curves for FIB1::monoGFP droplets at 30 minutes (open circles), 2 hours (open squares), and 4 hours (cross) after phase separation was initiated. The right plot shows fraction recovery of FRAP experiment as a function of time after phase separation for FIB1::GFP (closed circle) and FIB1::monoGFP (open circle) droplets. (C) Recovery time scale for *in vitro* (left) and *in vivo* (middle and right) FRAP experiments.

5.6 In vitro FIB1 and NPM1 Coexist as Multiphase Droplets

Given that FIB1 and NPM1 individually phase separate into liquid droplets in the presence of rRNA, we next tested how these proteins behave when mixed together. At relatively low concentrations, both proteins colocalize in the same condensed droplets; depending on the relative amount of FIB1 to NPM1, these droplets are either enriched in FIB1 (FIB1-rich/NPM1-lean phase) (Fig. 5.7B, 5.8B) or they are enriched in NPM1 (FIB1-lean/ NPM1-rich phase) (Fig. 5.7C, 5.8C). Thus, considering the soluble "buffer" phase, in these cases the system still resides

within a two-phase region of the phase diagram (Fig. 5.7D). However, when these proteins are both mixed at relatively high concentrations, we observe a three-phase system with both FIB1-rich/NPM1-lean droplets coexisting with FIB1-lean/NPM1-rich droplets, surrounded by the buffer phase (Fig. 5.7A). Interestingly, the NPM1 rich phase tends to partially envelope the FIB1 rich phase (Fig. 5.8A); in the absence of the GFP tag, this envelopment becomes even more pronounced, with FIB1 droplets fully embedded within NPM1 (Fig. 5.8D). This droplet organization is very similar to what is observed in *X. laevis* and mammalian nucleoli, where the FIB1 DFC is always internalized within the NPM1 GC. A phase diagram can be constructed by determining the threshold concentrations of FIB1 and NPM1 required to phase separate into a three-phase system (Fig. 5.7D). These results suggest that the "layered" structural organization of nucleolar proteins could be self-organized by liquid-liquid phase separation alone.

To test whether qualitatively similar phase behavior may be observed upon changing the relative protein concentrations in living cells, we injected nucleolar proteins into living *X. laevis* nuclei (Fig. 5.7E). Consistent with the expectation from equilibrium phase coexistence, we observed that the volume fraction of the corresponding component increased after microinjection (Fig. 5.7H). Typically, the DFC visualized by FIB1 is $25 \pm 2\%$ (mean ± s.e.m.) of the volume in the nucleolus. When more FIB1 was injected, the protein localized preferentially to the DFC causing the fibrillar cores to increase in size, occupying about $37 \pm 3\%$ (mean ± s.e.m.) of the volume. Conversely, when NPM1 was injected, the protein localized preferentially to the GC and caused the nucleoli to swell to large sizes, causing the fibrillar cores to occupy a lower volume fraction of only $15 \pm 1\%$ (mean ± s.e.m.). Moreover, small extranucleolar droplets of the respective protein appeared to form *de novo* (Fig. 5.7F,G). This is possible if the saturation

concentration in the nucleoplasm has been reached, causing spontaneous condensation of nucleolar proteins, without requiring NORs for nucleation [9].

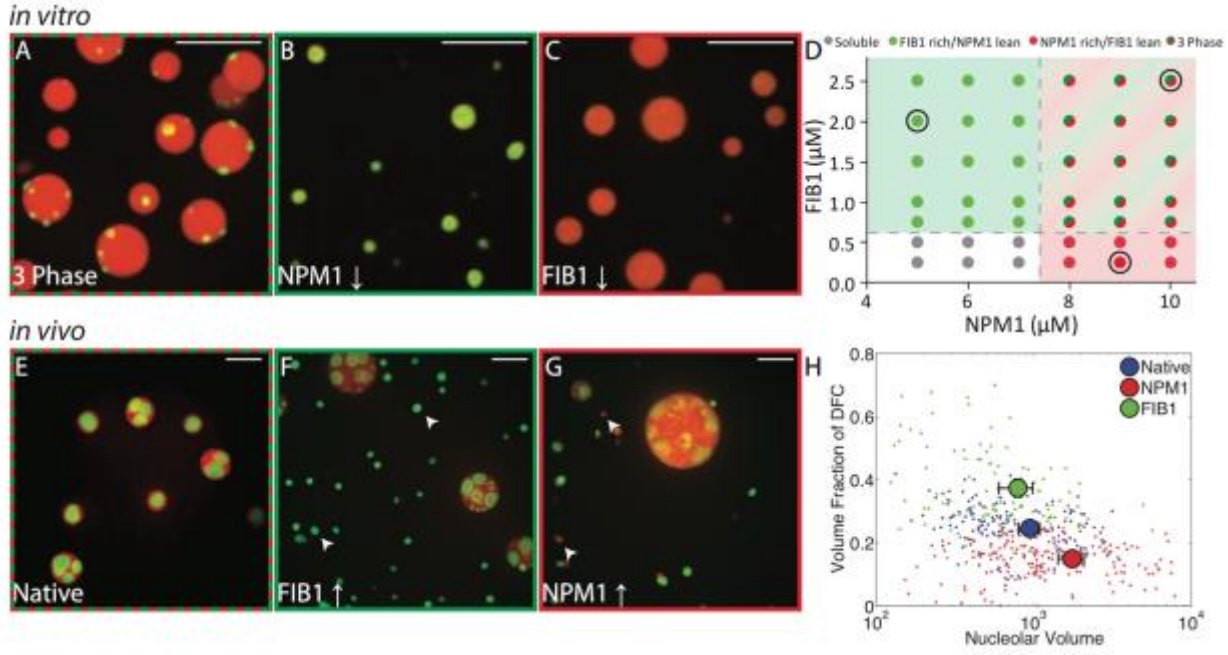


Figure 5.7: FIB1 and NPM1 form immiscible droplets *in vitro* and *in vivo.* (A-C) *In vitro* images of mixtures of purified NPM1 and FIB1. Scale bar = 10 μm. (A) High concentrations of both proteins (FIB1: 2.5 μM, NPM1: 10 μM) give rise to FIB1-rich droplets (green) which are immiscible with and partially enveloped by NPM1-rich droplets (red). (B) For much lower concentrations of NPM1 (NPM1: 5 μM, FIB1: 2 μM) only FIB1-rich/NPM1-lean droplets are observed. (C) For much lower concentrations of FIB1 (FIB1: 0.25 μM, NPM: 9 μM) only NPM1-rich/FIB1-lean droplets are observed. (D) Phase diagram for varying concentrations of NPM1 and FIB1 *in vitro*. Colors represent observed phase (gray = soluble phase, green = FIB1 rich/NPM1 lean phase, red = NPM1 rich/FIB1 lean phase, and red/green = three phase). Black circles indicate concentrations shown in A, B, and C. (E-G) Images of nucleoli in *X. laevis;* red=NPM1, green=FIB1. Scale bar = 10 μm. (E) Untreated nuclei. (F) Nuclei after microinjection of FIB1 (G) Nuclei after microinjection of NPM1. (H) Volume fraction of the DFC (identified by FIB1) in each nucleolus for native nuclei (blue), after NPM1 injection (red) and after FIB1 injection (green). Large symbols represent mean ± s.e.m.

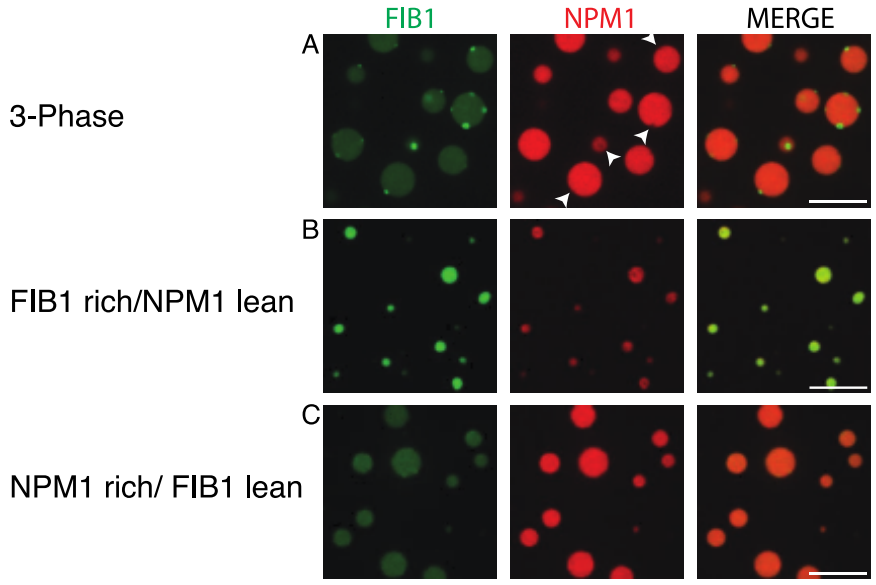


Figure 5.8: FIB1 and NPM1 form immiscible droplets *in vitro* and *in vivo.* (A) Single channel (left and middle) and merged (right) images of *in vitro* FIB1 (green) and NPM1 (red) three phase droplets. (B) Single channel (left and middle) and merged (right) images of *in vitro* FIB1-rich/NPM1-lean phase droplets. (C) Single channel (left and middle) and merged (right) images of *in vitro* FIB1-lean/NPM1-rich phase droplets. (D) Merged XY view (left), merged XZ view (bottom right), and single channel XZ view (top right and middle right) of *in vitro* FIB1-no tag (green) and NPM1 (red) three phase. Trace amount of FIB1::GFP used to visualize FIB1 droplets. Scale bar = 10 µm. Below, merged XY view of a single slice accompanied by a magnified version of a region indicated by white box.

5.7 Protein Domains Required for Phase Separation and Immiscibility

To gain insight into the molecular-scale driving forces underlying phase separation and droplet immiscibility, we created deletion mutants of both FIB1 and NPM1 that contained individual domains. Full-length FIB1 consists of two domains: a disordered N-terminal arginine (R)/glycine (G)-rich domain with low-sequence complexity (R/G domain) and an RNA methyltransferase domain (MD) that together with small nucleolar RNA (snoRNA) can methylate substrate rRNA (Fig. 5.9A). We find that the R/G domain (FIB1ΔC) is sufficient to form liquid-like droplets *in vitro*, while the MD alone (FIB1ΔN) is unable to form droplets *in vitro* (Fig. 5.9A). Interestingly, FIB1ΔC can phase separate into liquid-like droplets *in vitro,* even

in the absence of RNA (Fig. 5.4E). By contrast, full length FIB1 requires rRNA; however this may be a non-specific consequence of the polyanionic nature of rRNA since heparin can also drive phase separation of full length FIB1 (Fig. 5.4A,F). Furthermore, we find that FIB1ΔC droplets undergo homotypic fusion with an inverse capillary velocity of $\eta/\gamma = 0.5 \pm 0.04$ s/μm (95% confidence interval); these dynamics are significantly faster than for full length FIB1 ($\eta/\gamma \approx 40 \pm 10$ s/μm), and comparable to NPM1 ($\eta/\gamma \approx 0.3 \pm 0.07$ s/μm) (Fig. 5.3C,D, 5.4C). Consistent with this, in microrheology experiments, FIB1ΔC droplets also exhibit faster dynamics than full length FIB1 (blue symbols, Fig. 5.3E, 5.4C).

When these mutant proteins were injected into *X. laevis* nuclei, FIB1ΔN strongly partitions to the DFC, similar to the full-length FIB1 protein (Fig. 5.9C). Similarly, we observed that FIB1ΔN does not colocalize with NPM1 droplets *in vitro* (Fig. 5.9C, 5.10A); DFC enrichment of FIB1ΔN *in vivo* may not reflect immiscibility, but may instead reflect co-recruitment due to the presence of full length FIB1 in the native system. Moreover, we observed that injected FIB1ΔC localizes to the entire nucleolus, and has nonspecific interactions with the GC and DFC. Consistent with this, we find that the FIB1ΔC colocalizes with *in vitro* NPM1 droplets, rather than forming a third immiscible droplet phase (Fig. 5.10A). Taken together, these data suggest that the N-terminal R/G domain of FIB1 is sufficient for droplet formation, but does not encode for a separate liquid-like DFC sub-compartment; instead, the C-terminal MD of FIB1, which alone is not sufficient for droplet formation, confers immiscibility with proteins in the GC.

We next probed the importance of the three domains of NPM1: an N-terminal oligomerization domain (OD) which has been shown to be necessary to form an ordered pentameric structure [29], a central disordered domain containing acidic tracts (A2/A3), and a C-

terminal RNA binding domain (RRM) (Fig. 5.9B). Furthermore, the OD of NPM1 can form a pentamer to generate multivalency and could potentially increase the affinity of its RRM domain for rRNA. We deleted the N-terminal oligomerization domain to create NPM1ΔN, and we deleted the C-terminal RNA binding domain to create NPM1ΔC. We find that neither mutant is able to form droplets *in vitro*, consistent with phase separation of NPM1 requiring the oligomerization of NPM1 into multivalent pentamers that can bind to rRNA [4].

When NPM1ΔN is injected into *X. laevis* nuclei, we find that it localizes only very weakly to the nucleolus (Fig. 5.9D). This is consistent with the finding that pentameric state of NPM1 is necessary to retain this protein in the nucleolus [4]. When NPM1ΔN is mixed with FIB1 droplets *in vitro*, we see strong co-localization of NPM1ΔN within FIB1 droplets (Fig. 5.9D, 5.10A). To determine whether this strong co-localization is associated with the presence of rRNA in FIB1 droplets, we used poly-U50 to drive the phase-separation of FIB1. Interestingly, we find that the localization of NPM1ΔN within FIB1 droplets is reduced significantly with poly-U50 (Fig. 5.10B)*;* this suggests that the NPM1ΔN can localize within FIB1 droplets through its RRM interacting with rRNA. When NPM1ΔC is injected into *X. laevis* oocytes, we see that the protein strongly localizes to both the GC and DFC. However, NPM1ΔC localizes very weakly to FIB1 droplets *in vitro* (Fig. 5.9D, 5.10A)*;* this *in vitro* co-localization does not appear to be affected by the types of RNA used, suggesting that the interaction between NPM1ΔC and FIB1 *in vivo* is not driven by RNA, since NPM1ΔC lacks an RRM (Fig. 5.10C), but rather by interactions between the R/G domain of FIB1 and the OD/A2/A3 domains of NPM1 [29].

In summary, our domain analysis supports three key conclusions: 1) the disordered R/G domain of FIB1 can drive phase separation, but the time-dependent viscoelastic properties of

full-length FIB1 require the RNA-binding MD; 2) the disordered domains of both FIB1 and NPM1 appear capable of localizing equally to both components of nucleoli; and 3) the RNA binding domains, multivalent in the case of NPM1, play a key role in driving each protein to their respective sub-compartment.

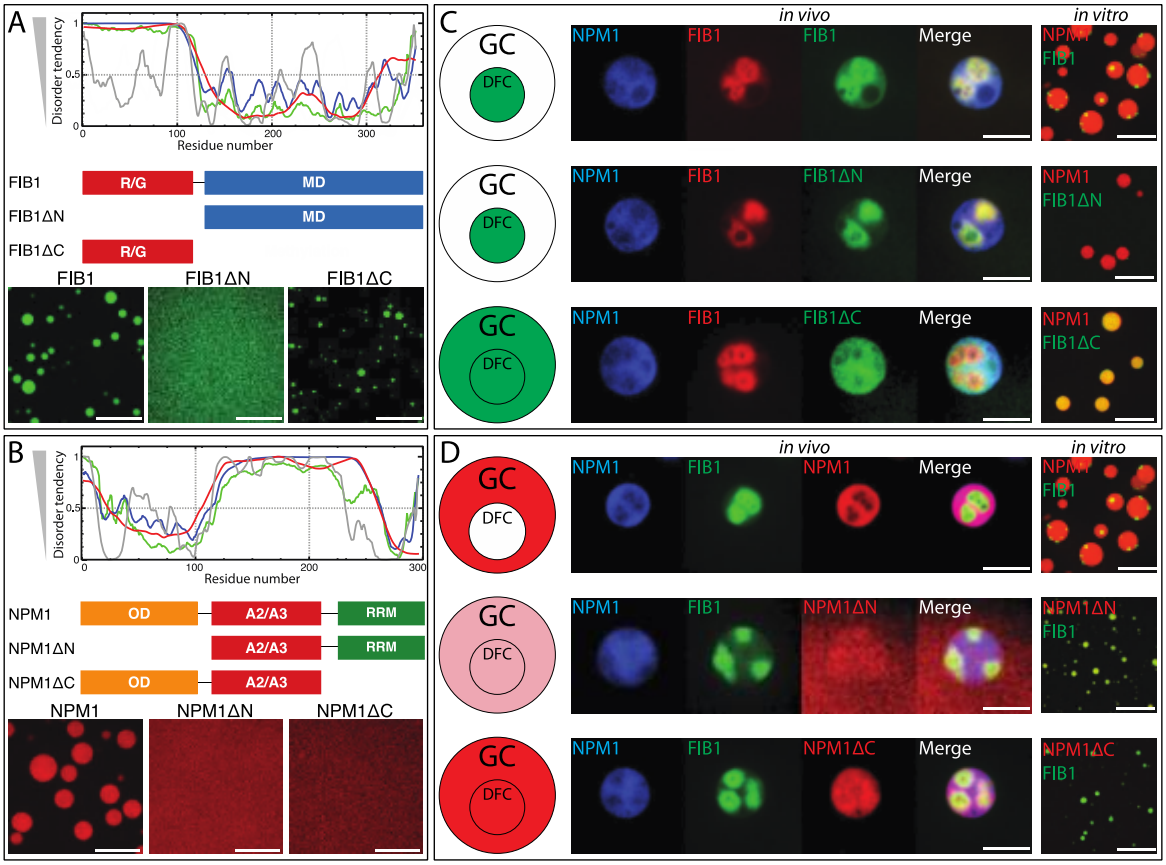


Figure 5.9: Molecular mechanism of phase partitioning in *X. laevis* oocytes. (A) Domain analysis of FIB1. Plot shows predicted disorder across full length FIB1 using various algorithms, P-FIT (green line), VSL2B (blue line), VL3 (red line), and VLXT (grey line). Schematic diagrams show three constructs: FIB1 full length, R/G deletion (FIB1ΔN), and deletion of MD (FIB1ΔC), with images below testing for constructs' ability to form droplets. Scale bar = 10 µm. (B) Domain analysis of NPM1. Predicted disorder across full length NPM1 for the four algorithms. Schematic diagrams show three constructs: NPM1 full length, oligomerization deletion (NPM1ΔN), and RNA binding deletion (NPM1ΔC) with images below testing for constructs' ability to form droplets. Scale bar = 10 µm. (C,D) The left most panel shows schematic summary of center panels. Center panels contain images from *X. laevis* nucleoli *in vivo*. Left channel contains expression of mRNA for NPM1::Cerulean, followed by expression of mRNA for FIB1::RFP or GFP, followed by injection of various constructs (FIB1, FIB1ΔN, FIB1ΔC, NPM1, NPM1ΔN, NPM1ΔC), and final image is the overlay of all three channels. Scale bar = 10 µm. The right most panel shows *in vitro* images of FIB1 or mutants (green) mixed with NPM1 droplets (red) or NPM1 or mutants (red) mixed with FIB1 droplets (green). Scale bar = 10 µm.

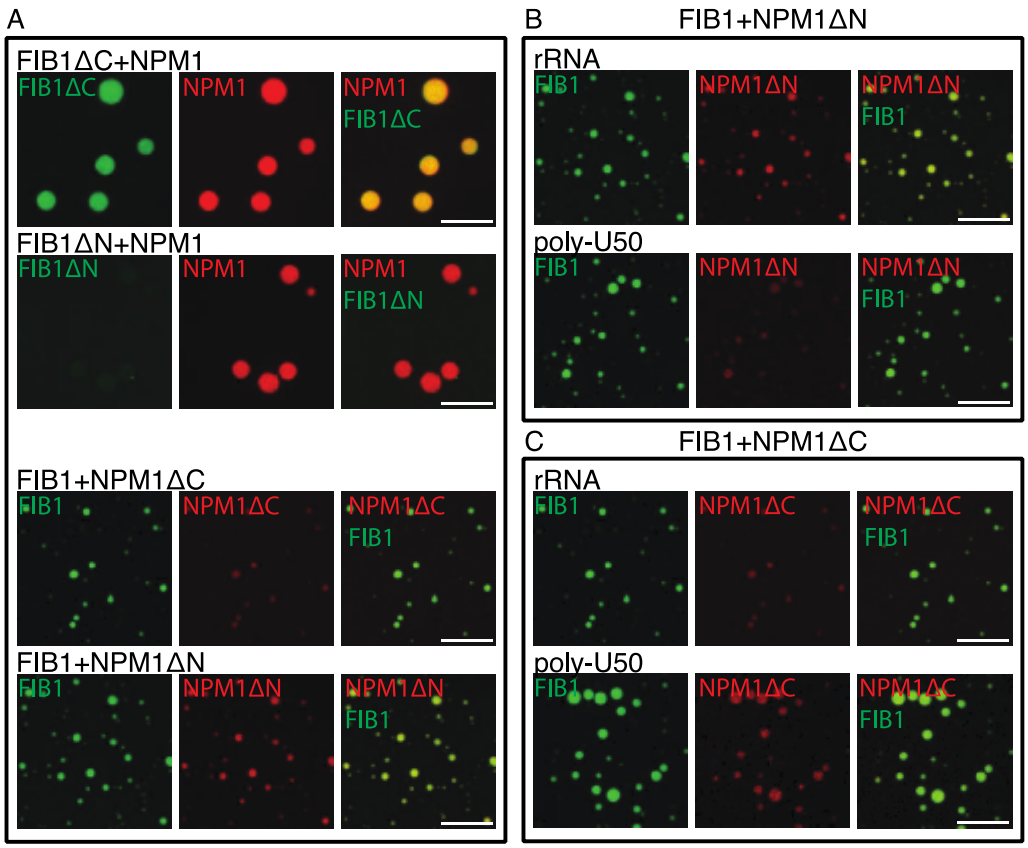


Figure 5.10: Molecular mechanism of phase partitioning in *X. laevis* oocytes. (A) The first two rows show single channel images (left and middle) and merged (right) images of *in vitro* NPM1 droplets (red) mixed with FIB1 mutants (green). The last two rows show single channel images (left and middle) and merged (right) images of *in vitro* FIB1 droplets (green) mixed with NPM1 mutants (red). Scale bar = 10 μm. (B) Localization of NPM1ΔN (red) in FIB1 droplets (green) in the presence of either rRNA (top row) or poly-U50 (bottom row). Scale bar = 10 μm. (C) Localization of NPM1ΔC (red) in FIB1 droplets (green) in the presence of either rRNA (top row) or poly-U50 (bottom row). Scale bar = 10 μm.

5.8 A Minimalist Computational Model for Three-Phase Behavior

Our data lead to the hypothesis that spatial organization within the nucleolus derives from the sequence-encoded interaction preferences of the different domains of nucleolar proteins. To test this hypothesis, we asked if the observed spatial organization could be reproduced in coarse-grained computer simulations. The simulation is comprised of 900 of each of the three polymers, performed on three-dimensional lattices to reduce the computational complexity. FIB1 and

rRNA were modeled as linear polymers of interaction modules, while the pentameric nature of NPM1 was captured using a branched polymer with five arms, each comprising the appropriate number of interaction modules (Fig. 5.11A).

Interactions between modules of each of the protein- and RNA-like polymers are governed by parameters of an interaction matrix. These parameters represent effective pairwise affinities in the presence of the competing effects of module-solvent and module-module interactions (Fig. 5.11B). The interaction matrix in figure 5.11B is sufficient to reproduce the totality of experimental observations. Figure 5.11C shows the normalized density profiles of FIB1, NPM1, and rRNA within droplets that form in the simulations, revealing a FIB1-rich core and NPM1-rich outer shell, with rRNA distributed across the two regions; figure 5.11D shows a representative cutaway snapshot from the simulations. Numerous distinct matrix parameterizations fail to reproduce one or more aspects of the *in vitro* data, although there are other specific choices of matrix parameters that do reproduce all of the experimental data (Fig. 5.12).

An exploration in the space of interaction matrix parameters suggests that the computational model must include three necessary features in order to reproduce all of the *in vitro* data. First, the R/G modules must have favorable homotypic interactions. Second, the OD of NPM1 should generate the requisite multivalency of RRM modules that drives the phase separation of NPM1 through interactions with rRNA. Third, the A2/A3 modules of NPM1 must be preferentially solvated, thus ensuring that they form weak or no bonds (Fig. 5.11B and 5.12L).

The minimalist model supports the presence of three distinct phases: Phase 1 is the solvent and includes water plus the solution ions; Phase 2 is NPM1 + rRNA; and Phase 3 is FIB1

+ rRNA. The balance of interactions can be quantified in terms of pairwise interaction coefficients designated as χ_{12} (Solvent-NPM1), χ_{13} (Solvent-FIB1) and χ_{23} (NPM1-FIB1) that are derived from the Flory-Huggins theory for polymer solutions and blends [30]. The individual χ values quantify the free energy gained or lost when modules exchange homotypic interactions for heterotypic ones. A direct consequence of the structure of the interaction matrix (Fig. 5.11B) is that the values of each of χ_{12}, χ_{13}, and χ_{23} are positive. The three-phase behavior observed in experiments and reproduced in simulations implies that χ values must obey the relation: $\chi_{13} > \chi_{12} > \chi_{23} > 0$. Since surface tension is directly proportional to the Flory parameter, $\gamma \sim \chi$, it follows that $\gamma_{13} > \gamma_{12} > \gamma_{23}$ i.e., the surface tension of FIB1 droplets is larger than that of NPM1 droplets. It is thus energetically more favorable to envelope the FIB1 droplet within the NPM1 droplet, as opposed to NPM1 being enveloped within a FIB1 droplet (Fig. 5.13D).

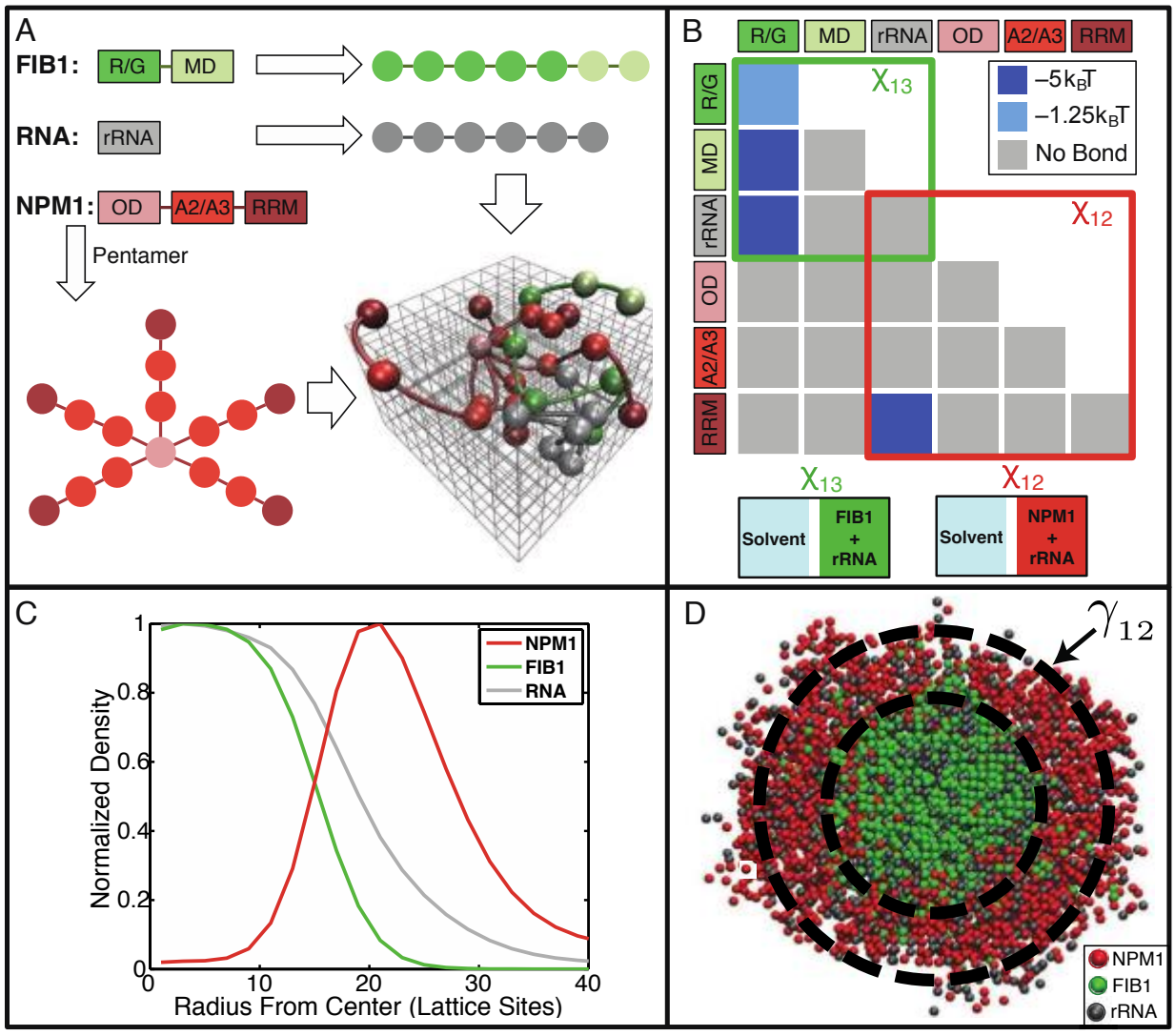


Figure 5.11: Preferential interaction model captures the formation of spatially organized droplets for the ternary system comprising of FIB1, NPM1, and rRNA. (A) Mapping of the sequences of FIB1, NPM1, and rRNA to linear/branched polymers of modules on three-dimensional lattices. FIB1 is modeled as a linear polymer comprising of seven modules, five corresponding to the R/G domain and two corresponding to the MD. Similarly, the rRNA sequence is modeled as a linear polymer comprising six modules. NPM1 is modeled as a branched polymer with five arms. Here, the ODs of five NPM1 molecules occupy the base for each branch; two modules correspond to the intrinsically disordered acid-rich regions (A2/A3) and a single module captures the RNA recognition module (RRM). A representative snapshot is shown of polymers on the cubic lattice. (B) The matrix of module interaction strength. (C) The normalized mean radial density of FIB1 (green), NPM1 (red), and RNA (grey) for representative largest cluster observed throughout a simulation. (D) Visual depiction of a slice through representative phase separated droplet; FIB1 (green) and NPM1 (red).

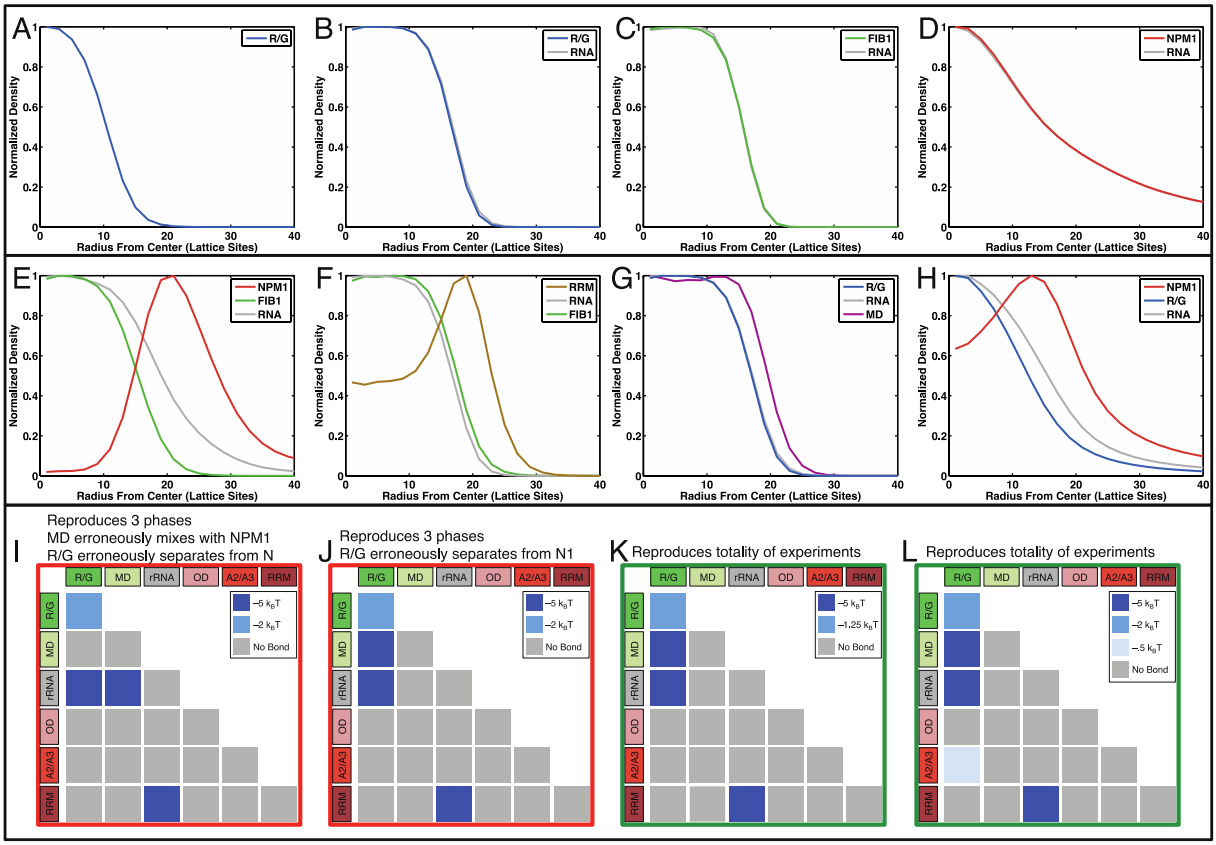


Figure 5.12: Preferential interaction model captures the formation of spatially organized droplets for the ternary system comprising of FIB1, NPM1, and rRNA. (A) The radial density profile showing that the interaction matrix in figure 5.11B leads to uniform, spherical droplets of the R/G domain. (B) The phase separation driven by R/G is reproduced in the presence of rRNA and the latter are well mixed with R/G molecules in the droplet thus leading to nearly superimposable density profiles. (C) Full length FIB1 makes liquid-like droplets in the presence of rRNA. (D) Radial density profile provided as proof that NPM1+rRNA phase separates to form droplets. (E) This figure is identical to figure. 5.11C and is reproduced here for completeness and to provide a visual guide for interpreting the results in panels F-H. (F) Colocalization of the RRMs of NPM1 into FIB1 droplets. (G) Colocalization of the MD of FIB1 into droplets formed by R/G and rRNA. (H) Colocalization of R/G domains into droplets formed by NPM1+rRNA. In panels B-H the legend designates rRNA as RNA. (I) and (J) Examples of interaction matrices that reproduce three-phase behavior for the ternary system of NPM1, rRNA, and FIB1 but fail to reproduce the experimental observations for truncation constructs. (K) – (L) Examples of interaction matrices that reproduce the totality of experimental observations. Panel K is identical to figure 5.11B.

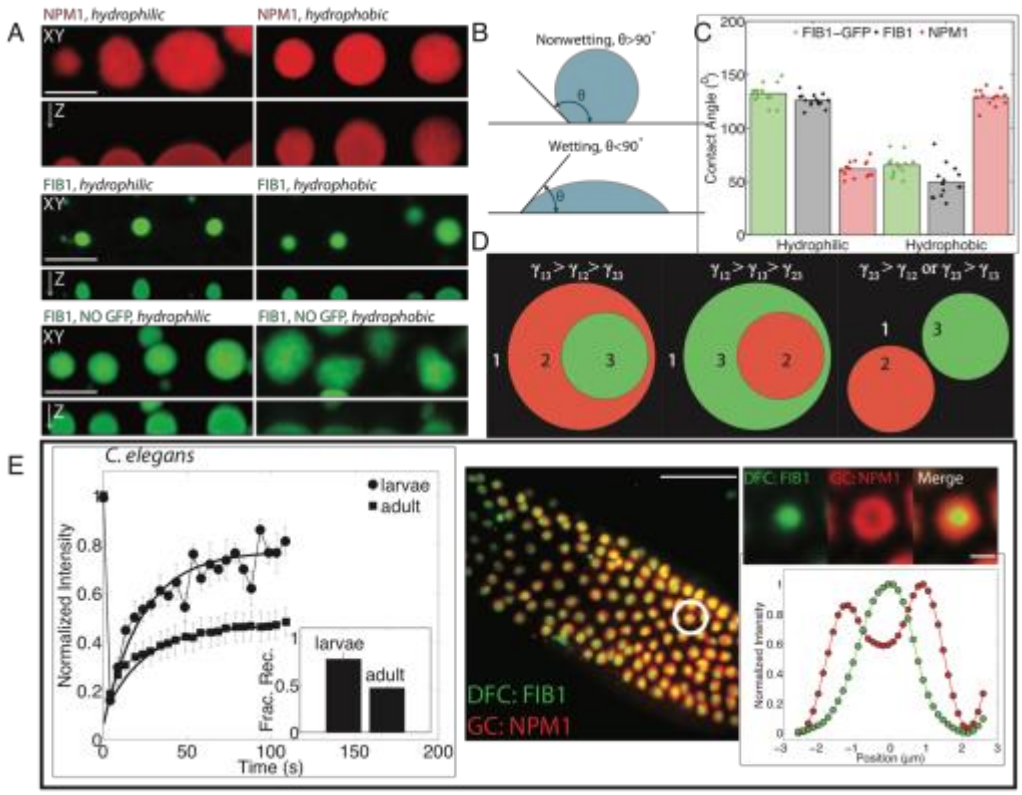


Figure 5.13: Surface tension drives organization of multiphase droplets. (A) Top row contains XY projections of *in vitro* droplets (NPM1, FIB1::GFP, and FIB1-no tag) on either pluronic (hydrophilic, left) or sigma-cote (hydrophobic, right) treated coverslips. Bottom panel contains XZ projections of *in vitro* droplets. Scale bar = 5 μm. (B) Schematic diagram showing side view of a drop with a contact angle ϑ>90° under non-wetting conditions (top) or with a contact angle ϑ<90° wetting conditions (bottom). (C) Contact angles measured for FIB1::GFP (green), FIB1-no tag (grey), or NPM1 (red), measured on pluronic (hydrophilic) or Sigmacote (hydrophobic) treated coverslips for 15 droplets under each condition. (D) Schematic diagrams showing localization of different phases depending on relative surface tensions. (E) FRAP recovery curves of FIB1:GFP in nucleoli of *C. elegans*. Circles are traces for larvae and squares are traces for adult worms. Inset shows fraction recovery for larvae and adults. (F) Images of the *C. elegans* gonad expressing FIB1::GFP (green) and NPM1::mCherry (red). Enlarged images of the green, red and merged channels of circled nucleolus. Intensity profiles of FIB1::GFP (green) and NPM1::mCherry (red) for a line drawn through the center of the nucleolus.

5.9 Nucleolar Organization Arises from Differential Surface Tension of Sub-Compartments

The physical picture that emerges from our computational model is consistent with the very low values we obtained for the surface tension of the *in vitro* NPM1 droplets, as well as for the *in vivo* NPM1-rich GC. Unfortunately, the small size of *in vitro* FIB1 droplets, as well as their viscoelasticity, makes it difficult to undertake direct surface tension measurements. As an

alternate route to evaluate the relative droplet surface tensions, we measured droplet wetting behavior on hydrophobic and hydrophilic coverslips. Wetting refers to the contact between liquids and surfaces and is a consequence of surface tension; for example, water droplets will spread over a Pluronic-treated hydrophilic surface (low contact angle), while water droplets will round up and avoid contact with a Sigmacote-treated hydrophobic surface (high contact angle), as shown in figure 5.14 A,B.

On hydrophobic surfaces, we find that NPM1 droplets behave as water droplets and exhibit minimal wetting, with a contact angle of 130±10° (mean±s.d) (Fig. 5.14E, 5.13A,C). On hydrophilic surfaces they exhibit increased wetting, with a contact angle of 60±10° (Fig. 5.14C, 5.13A,C). In contrast, FIB1 droplets tended to better wet the hydrophobic coverslips, with a contact angle of 70±10° (mean±s.d) (Fig. 5.14F, 5.13A,C), as compared to the hydrophilic coverslips, on which they exhibited a contact angle of 130±10° (Fig. 5.14D, 5.13A,C). The differential hydrophobicity of NPM1 and FIB1 droplets explains our key observation, which we describe *in vivo* (Fig. 5.1 A-G), *in vitro* (Fig. 5.7A), and also *in silico* (Fig. 5.11C-D): FIB1 and NPM1 form multiphase droplets where FIB1 is at least partially encapsulated by NPM1 (Fig. 5.14I). This organization is quite similar to how immiscible liquids are organized in non-biological multiphase systems [31]. To demonstrate this with a simple example, we prepared a system of water, Crisco vegetable oil, and silicone oil, which are immiscible liquids with known surface tensions [24] (Fig. 5.14G). Silicone oil is more hydrophobic than Crisco oil, i.e. $\gamma_{silicone/water} > \gamma_{Crisco/water}$, and as a result, the silicone oil droplets are always enveloped by the Crisco oil droplet. Similarly, since FIB1 is more hydrophobic than NPM1, $\gamma_{FIB1/water} > \gamma_{NPM1/water}$, FIB1 droplets will tend to be encapsulated within NPM1 droplets (Fig. 5.14H, 5.13D). We note

that in both cases there is also the requirement that a third surface tension, $\gamma_{silicone/Crisco}$ or $\gamma_{FIB1/NPM1}$, must not be too high, or the two droplets would never interact (Fig 5.13D).

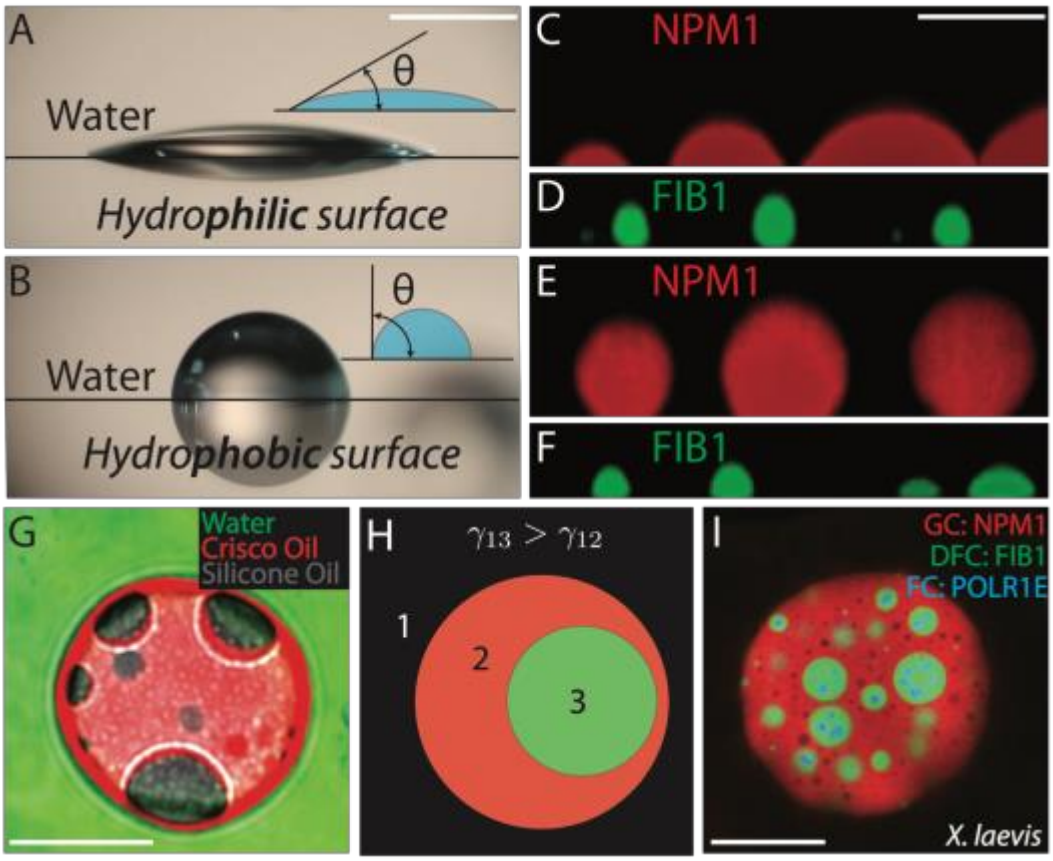


Figure 5.14: Surface tension drives organization of multiphase droplets. (A-F). Images of droplets on hydrophilic surfaces (Pluronic-treated, A) or hydrophobic surfaces (Sigmacote-treated, B). (A) Water droplet on hydrophilic surface. (B) Water droplets on hydrophobic surface. Scale bar for A, B = 1 mm. (C) NPM1 droplets on hydrophilic surface. (D) FIB1 droplets on hydrophilic surface. (E) NPM1 droplets on hydrophobic surface. (F) FIB1 droplets on hydrophobic surface. Scale bar for C-F= 5 μm. (G) Image of non-biological multiphase droplets: green=water, red=Crisco oil, and gray=silicone oil. Scale bar = 20 μm. (H) Schematic organization of immiscible multiphase droplets. The more hydrophobic phase (green), has a higher surface tension with water than the more hydrophilic phase (red), which has a lower surface tension with water. (I) Image of multiphase nucleoli after actin disruption in *X. laevis*. Scale bar = 20 μm.

5.10 Discussion

The nucleolus is the most prominent of numerous membrane-less RNP bodies and was recognized over 150 years ago by early microscopists [32]. However, a mechanistic biophysical understanding of the principles governing the well-known sub-compartmental organization has been elusive [33]. Our findings suggest that these distinct compartments arise as a consequence

of liquid phase immiscibility, supported by: 1) The different layers can undergo coalescence upon contact, relaxing toward round droplet-like structures; 2) Protein components of the different layers are highly dynamic as determined by FRAP; 3) When expressed and purified, key enriched nucleolar proteins undergo phase separation into droplets with properties comparable to those found in their corresponding nucleolar sub-compartment; 4) When mixed, purified proteins exhibit phase immiscibility, resulting in sub-compartmentalized *in vitro* droplets that are strikingly similar to *in vivo* nucleoli; 5) A minimalist coarse-grained model is sufficient for reproducing phase immiscibility and the observed colocalization of different truncation constructs, and further shows how a network of molecular interactions determines surface tensions, which dictate the core-shell droplet architecture; and 6) Biophysical characterization and droplet wetting behavior indicate that the FIB1/DFC phase exhibits a higher surface tension than the NPM1/GC phase, which results in the former being embedded within the latter.

What are the functional implications of liquid phase immiscibility and surface-tension driven sub-compartmentalization of the nucleolus? The most important nucleolar function is ribosome biogenesis [18], which involves the coordinated modification and assembly of rRNA through multiple processing steps [34, 35]. Analogous to an assembly line or the staged processes in a modern chemical plant, the spatial separation and distinct physical and compositional features of the FIB1-rich DFC and NPM1-rich GC may tune the vectorial transport and associated processing of rRNA into mature pre-ribosomal particles. Indeed, continuous transcription within the FC causes radial flux of rRNA through the DFC into the GC and finally into the nucleoplasm. The DFC is effectively an enzymatic bath, which facilitates correct base-pairing with small nucleolar guide RNAs (snoRNAs), for example those associated

with methylation by FIB1 as part of a box C/D snoRNP and pseudouridinylation by H/ACA snoRNPs, as well as cleavage reactions and other rRNA modifications [35]. These modifications are critical for correct rRNA folding and stability, subsequent assembly with ribosomal proteins, and ultimately ribosome function (*i.e.,* translational fidelity) [36].

We propose that the viscoelastic properties of the DFC serve to lower the flux of incompletely or incorrectly processed/folded pre-ribosomal particles, ensuring that DFC-associated enzymatic processes are completed, before passage of pre-ribosomal particles into the outer GC layer, where they encounter NPM1 and early binding ribosomal proteins. Indeed, NPM1 phase separates with both rRNA and ribosomal proteins [4], consistent with fluorescence imaging studies suggesting that ribosomal proteins localize to the GC, but not the DFC [37]. Our findings indicate that pentameric NPM1 is integral to the fluid features of the GC, whose relatively low viscosity may allow ribosomal proteins to remain dynamically accessible to pre-ribosomal particles emerging from the DFC.

Our data show that FIB1 droplets, but not NPM1 droplets, are metastable and can age with time both *in vitro* and *in vivo.* These data are consistent with the hypothesis that disordered regions can facilitate the transition from liquid-like to solid-like structures [38], which is supported by recent studies on a variety of RNA binding proteins [11, 12, 26, 39, 40]. We note that the name "Fibrillarin" was given due to its localization to fibrillar structures within the DFC [41]; these structures could reflect droplet aging/fibril formation. Consistent with this, we observe apparent aging of nucleoli in non-dividing *C. elegans* intestinal cells: FIB1 recovers less completely in older adult worms compared with younger larvae (Fig. 5.13E). However, given that RNA can impact the fluidity of related phase separated droplets [13, 26], the rate of FIB1-rich DFC aging could be impacted by the flux of newly synthesized rRNA transiting through the

DFC. Future studies will be required to elucidate the biophysical origin of nucleolar droplet maturation and the role of RNA transcription and other ATP-dependent processes.

Our findings underscore the importance of surface tension, whose role is well established in physical systems, as is readily visualized with immiscible oils in water (Fig. 5.14G). In living systems, effective surface tension may be important for the organized demixing of cell populations: cell types that exhibit a relatively high apparent surface tension will tend to be enveloped by cell types with a relatively low apparent surface tension [42]. Our data show that this same basic principle is important for structuring the nucleolus, with possible implications for other RNP bodies. For example, histone locus bodies (HLBs) in the frog nucleus are commonly found to have B-snurposomes attached to their surface [43]; incomplete internalization of B-snurposomes suggests that their surface tension may be similar to HLBs. Interestingly, this partial internalization is reminiscent of the altered nucleolar structure observed in actinomycin-D treated nucleoli, wherein rRNA transcriptional inhibition results in a more lobulated nucleolus [44, 45]. A similar organization is also seen with processing bodies [20, 46, 47]. Recently, stress granules have been shown to contain less dynamic cores, which exhibit a qualitative similarity to the FIB1/DFC cores of the nucleolus [19]. Building on the biophysical groundwork we have laid here, it may be possible to alter or even invert the organization of such RNP bodies, by using surfactants to modulate droplet surface tensions; this could significantly impact sequential RNA processing steps and the overall flow of genetic information.

Organelle sub-compartmentalization is well-known in membrane-bound organelles, such as mitochondria. Our data show that membrane-less liquid phase organelles can also generate significant substructure. Phase separation and the coexistence of multiple distinct liquid

RNA/protein phases thus provide a simple but elegant mechanism for the cell to control the spatial localization and processing of molecules, without relying on membrane boundaries.

5.11 Experimental Procedures

Preparation of *X. laevis*, mammalian, and *C. elegans* nucleoli. Frogs were anesthetized with 0.1% MS-222 solution, and oocytes were surgically removed from female *X. laevis* frogs following an IACUC approved protocol. mRNA of endogenous proteins (FIB1, NPM1, and POLR1E) with fluorescent tags and recombinant proteins were microinjected into oocytes. Nuclei were manually dissected in mineral oil and subsequently imaged. Actin was disrupted using Lat-A, and ATP was depleted using Apyrase. Mammalian cells expressing fluorescent fusion proteins (FIB1 and NPM1) were maintained at 37°C using standard conditions, and ATP was depleted using sodium azide and deoxyglucose. *C. elegans* expressing intestinal FIB1::GFP were maintained at 20°C under standard conditions and anesthetized with levamisole in M9 prior to imaging.

Purification and phase separation of *in vitro* droplets. FIB1 and NPM1 variants were expressed using a standard *E. coli* expression system, purified using either a 6x-His or GST tag, and stored in a high salt buffer. Phase separation was initiated by lowering the salt concentration of stock protein in the presence or absence of rRNA. Non-biological multiphase droplets were obtained by vortexing water, Crisco oil, and silicone oil.

Biophysical characterization of *in vivo* and *in vitro* droplets. For fusion relaxation experiments, the aspect ratio was measured as a function of time for droplets of different size to obtain the inverse capillary velocity. Surface tension of non-wetting droplets with measured densities was estimated from non-spherical XZ shape profiles obtained using a right-angle prism.

For microrheology experiments, time-lapse images of fluctuating R=50 nm particles inside protein droplets were acquired and analyzed using particle-tracking Matlab code to obtain the mean squared displacement as a function of lag time; from the Stokes-Einstein relation, the viscosity was determined. For FRAP experiments, 1 μm spots inside *in vivo* and *in vitro* droplets were photobleached, and percent fluorescent recovery and recovery times were determined using standard techniques. Wetting behavior of *in vitro* droplets was observed for surfaces treated with Sigmacote (hydrophobic) or Pluronic (hydrophilic), and the contact angle was measured at the interface between the glass and line tangent to the droplet.

***X. laevis* oocyte collection.** Frogs were anesthetized with 0.1% MS-222 solution for 15 minutes, and oocytes were surgically removed from adult female *X. laevis* frogs following an IACUC approved protocol as previously described [22]. Oocytes were incubated at 18°C in OR2 solution. To remove the follicular layer, the oocytes were first mechanically separated and then incubated for 1 hour and 20 minutes in 2 mg/ml collagenase (Sigma). Stage V-VI oocytes of diameter of 1-1.3 mm were used for all experiments and identified using a Zeiss stereoscope [48].

DNA and mRNA constructs for *X. laevis*. Vector pCS2+ backbones were used for all fluorescent fusion constructs. The granular component was visualized either with NPM1::GFP, NPM1::RFP, or NPM1::Cerulean; the dense fibrillar component was visualized with FIB1:: GFP or FIB1::RFP, and the fibrillar component was visualized with mCherry::POLR1E or GFP::POLR1E. The nuclear actin network was visualized with a Lifeact::GFP construct [22].

Purification of nucleolar proteins. Recombinant versions of FIB1::GFP protein with a N-terminal 6×-His tag and NPM1 with a N-terminal GST tag were purified using the *E. coli* expression system, BL21(DE3) cells. For FIB1::GFP, cells were lysed in resuspension buffer (20

mM Tris-HCl, pH 7.5, 500 mM NaCl, 10 mM imidazole, 14 mM β-mercaptoethanol, and 10% (vol/vol) glycerol) containing 1 mg/mL lysozyme and a protease inhibitor mixture (Roche Diagnostics). FIB1::GFP was captured with Ni-NTA agarose (Qiagen), washed well with Ni-Wash buffer (20 mM Tris·HCl, pH 7.5, 500 mM NaCl, 14 mM β-ME, 10% (vol/vol) glycerol, and 25 mM imidazole), and eluted with Ni-Elution buffer (20 mM Tris, pH 7.5, 500 mM NaCl, 14 mM β-mercaptoethanol, 10% (vol/vol) glycerol, and 250 mM imidazole). Furthermore, elution from Ni-NTA was loaded onto a HiTrap Heparin column (GE) after being diluted in heparin binding buffer (20 mM Tris, pH 7.5, 50 mM NaCl, 1% (vol/vol) glycerol, and 2 mM DTT) and eluted in 20 mM Tris, pH 7.4, 1 M NaCl, 1% (vol/vol) glycerol, and 2 mM DTT. Glycerol was added to 10% (vol/vol), and aliquots were flash frozen in liquid nitrogen and stored at −80 °C.

For NPM1, cells were lysed in resuspension buffer (20 mM Tris, 300 mM NaCl, 10 mM β-mercaptoethanol, protease inhibitors (Sigma-FAST), 1 mM EDTA, pH 7.5) containing Benzonase (Millipore, 20U/uL). GST-NPM1 was captured using GSH beads, washed well with wash buffer (20 mM Tris, 300 mM NaCl, 1 mM EDTA, pH 7.5), and eluted with elution buffer (20 mM Tris, 300 mM NaCl, 10 mM BME, 10 mM reduced L-glutathiol, 1 mM EDTA, pH 7.5). Eluted protein was dialyzed in the presence of Turbo3C/HRV3C/PreScission protease (Biovision, cat #. 9206-1) against 10 mM Tris, 0.15 M NaCl, 2 mM DTT, pH 7.5 overnight at 4C. Furthermore, HPLC was performed and eluent was lyophilized before storing in -20C.

Phase separation *in vitro.* For *in vitro* experiments, frozen FIB1 aliquots were thawed at room temperature and buffer exchanged (Amicon; 0.5 mL, 3–10k) into freshly made high salt buffer (20 mM Tris, pH 7.5, 1 M NaCl, and 1 mM DTT) to inhibit droplet formation. Similarly, lyophilized NPM1 was resuspended in Guanidine-HCl and refolded via dialysis in 20 mM Tris,

0.15 M NaCl, 2 mM DTT, pH 7.5 overnight. Protein solutions were subsequently mixed with high purity wheat germ rRNA (BioWorld) and varying volumes of salt buffer (20 mM Tris, pH 7.5, and 1 mM DTT with varying NaCl concentration) to obtain final rRNA concentrations of 5 μ g/mL for FIB1 and 100 μ g/mL for NPM1and desired protein/salt concentrations. NPM1 labeled with Dylight 594 NHS Ester (ThermoFisher Scientific) was added in trace amount to visualize NPM1 droplets in imaging-based assays. Samples were prepared in imaging chambers using silicone wells (Grace BioLabs) and observed under a microscope to score for phase behavior after incubation of 30 minutes onward. For three-phase assays, FIB1::GFP and NPM1 were phase separated with 5µg/mL and 100µg/mL of rRNA respectively at 150 mM NaCl buffer, mixed together after 5 minutes, and incubated for 30 minutes prior to imaging.

Effect of fluorescent tag on *in vitro* phase separation of FIB1 protein. Since the main FIB1 construct we work with is GFP tagged, we investigated whether this GFP tag alters the phase separation behavior and physical properties of the *in vitro* droplets. We explored the phase boundary and fusion dynamics of untagged FIB1 droplets (visualized with sparsely labeled RNA), as well as droplets of FIB1 tagged with a GFP mutant (A206K) which is known to exist as a monomer [49] (Fig. 5.4B). We find that the various constructs do exhibit somewhat different behavior. For example, the GFP-tag shifts the phase boundary, requiring roughly 2-fold less protein to phase separate as compared to the untagged GFP. The inverse capillary velocity of FIB1 is also impacted roughly 2-fold by the GFP tag (Fig. 5.4B). Nonetheless, the impact of various means of tagging proteins is insignificant compared to the more than 10-fold difference in properties of FIB1 versus NPM1 droplets.

Microinjection of mRNA & protein constructs and nuclei dissection. Nuclei were microinjected with a Narishige micromanipulator and PicoPump PV820 as previously described

[22]. mRNA constructs were microinjected into the cytoplasm, and oocytes were allowed to incubate overnight at 18C. Proteins were injected directly into the nucleus, and oocytes were allowed to incubate for at least 2 hours before imaging. Nucleolar proteins NPM1 conjugated with dylight and FIB1::GFP were microinjected directly into the nucleus at an initial concentration of 32 μM and 2 μM, respectively, in 150 mM NaCl, 10 mM Tris pH 7.5, and 20 mM DTT. Oocytes were allowed to recover and were subsequently imaged at least 2 hours after microinjection. For all experiments, nuclei were manually dissected using forceps and a hair loop in mineral oil under *in vivo* conditions [50].

Actin disruption and coarsening. To disrupt actin, latrunculin A (Lat-A, Sigma) treatment was performed for 1-2 hours at 2 μg/ml at constant rotation. After Lat-A treatment, nuclei were dissected in mineral oil and placed in an imaging chamber consisting of a glass coverslip and glass coverslide separated by a silicone well (Grace Biolabs) as previously described [22]. Movies capturing all three nucleolar compartments consisted of 10-15 μm z-stacks with 1-2 μm step size and were acquired for several hours. Maximum intensity projections were made in each channel, and fusion events were analyzed in time for the granular and dense fibrillar components. The aspect ratio of each nucleolar phase was determined as $A.R. = \ell_{long}/\ell_{short}$ as a function of time, where ℓ_{long} and ℓ_{short} represent the lengths of the long and short axes of the nucleolar phase. The data was fit to an exponential to determine the relaxation time: $A.R. = a + b \cdot \exp(-t/\tau_f)$, and the characteristic length scale, ℓ, of the nucleolar compartment was the measured radius. The time scale for fusion, τ_f, is expected to be directly proportional to the characteristic length scale, ℓ, of droplets according to the relation: $\tau_f \approx (\eta/\gamma) \cdot \ell$. Here, the inverse capillary velocity [21] is the ratio of the viscosity of the droplet, η, to surface tension, γ, which underlies the spherical shape of droplets.

Surface tension measurements in *X. laevis* nuclei. After actin disruption by Lat-A, dissected nuclei were placed in glass bottom petri dishes (MatTek) filled with mineral oil. A 0.5 mm right angle prism (Edmund Optics) was manually placed adjacent to nuclei to visualize the XZ dimension. For experiments of the steady-state shape profile, nuclei were imaged after 2-3 hours of incubation in the petri dish. For experiments involving the relaxation of nucleolar shape, the oocytes were left to sit overnight after actin disruption, so that the nucleoli had time to sediment and fuse into one massive nucleolus (R>15 μm). The next day, those oocytes were rotated to cause the massive nucleolus to round into a sphere for several hours. Finally, nuclei were dissected and immediately imaged to capture the rapid deformation under gravity.

Using a right angle prism imaging approach to avoid imaging artifacts along the optical axis (Z), we examined the steady-state shape profile in the XZ dimension (Fig. 5.1J). The XZ shape profile is a balance between surface tension, which will promote rounder droplets, and gravitational forces, F_g, (along the optical Z-axis), which will tend to flatten droplets. The shape profile of the brightest XZ frame was obtained from custom image analysis and two length scales were obtained: R, which is the distance from the center to the widest point and H, which is the height of the droplet (Fig. 5.3b inset). The surface tension was obtained as $\gamma = \Delta\rho g H^2/B \approx \Delta\rho g H^2/4.308[1 - H/R]$, where $\Delta\rho$ is the known density difference between the nucleolus and the surrounding nucleoplasm [22], g is the acceleration due to gravity, and B is the empirically determined shape factor, which is a function of H/R ratio [51].

The preceding analysis focuses on the steady state shape of large nucleoli deformed under the force of gravity. However, the timescale over which the nucleolus deforms to this shape can also yield insights into its properties. From dimensional analysis, the time scale for this shape relaxation is given by $\tau_g \approx (\eta/\gamma^2) \cdot F_g \approx (\eta/\gamma^2) \cdot \Delta\rho \cdot g \cdot \ell^3$, where g is the

gravitational acceleration. We measured relaxation times on the order of 10-30 minutes for large nucleoli of diameter >30 microns (Fig. 5.1L, Supplemental Video 4). Solving for viscosity, we obtain values of 30 ± 10 Pa·s (mean ± s.e.m). From these measurements of surface tension and viscosity, we obtain an independent estimate of the ratio of surface tension to viscosity: $\eta/\gamma \approx 50 \pm 10$ s/μm (mean ± s.e.m) (Fig. 5.1L inset). These measurements are consistent with those made from fusion relaxation experiments (Fig. 5.1H).

Surface tension measurements of nucleolar proteins *in vitro.* Purified nucleolar proteins (NPM1-dylight) were allowed to phase separate with RNA and were gently centrifuged to form large droplets. Small volumes from the phase separated solution were pipetted into a glass bottom dish filled with mineral oil (Sigma). Glass bottom dishes (MatTek) were previously treated with sigma cote (Sigma) for NPM1 droplets to create non-wetting conditions. XZ shape profiles were imaged by using a 0.5 mm prism (Edmund optics) and analyzed as described in the previous section. By analyzing the sedimentation rate of NPM1 droplets [22], we obtained a density difference between NPM1 droplets and the surrounding low concentration solution of $\Delta\rho = 60 \pm 20\ kg/m^3$, which was used along with the shape profile to estimate the surface tension as described above.

Wettability of protein droplets. Coverslips were treated with 1% Pluronic F-127 solution (Sigma Aldrich) to make the surface hydrophilic or with Sigmacote (Sigma Aldrich) to make the surface hydrophobic. Solutions of Pluronic or Sigmacote were placed on coverslips for approximately 10 min and washed off with DI water (Millipore) and dried with nitrogen gas. Protein droplets were placed in imaging chambers containing the treated coverslips. 3-D volume stacks were acquired and projected in XZ to obtain the shape profile. Contact angles were

measured in Image-J as the angle between the line tangent to the drop and contact line interior of the drop.

Microrheology of protein droplets. Microrheology was performed in FIB1 and NPM1 droplets by adding R=50 nm fluorescent polystyrene microspheres (Invitrogen) to protein solutions inside an imaging chamber. Using spinning disk confocal microscopy, time-lapse movies were acquired 1-2 μm above the coverslip with a 100 ms interval and an exposure time less than one-fifth of the acquisition interval. Images were analyzed using particle-tracking algorithms as previously described [22], and the two-dimensional mean-squared displacement was calculated as a function of lag-time. We fit the data to obtain the diffusive exponent and the diffusion coefficient; using the Stokes-Einstein equation, we obtained the viscosity. The noise floor was obtained by performing similar experiments on R=50 nm fluorescent polystyrene microspheres dried onto a glass coverslip.

Handling and imaging of mammalian cells. NIH 3T3 fibroblast cells were maintained in DMEM media supplemented with 10% FBS, 1% Penicillin/Streptomycin, and 1% Glutamax 100X. For maintenance, cells were trypsinized and passaged when they reached 70-80% confluence. For imaging, cells are plated on fibronectin-coated glass bottom dishes in HBSS/2% FBS and imaged using a 37°C heating stage. All images are taken with a Nikon A1 laser scanning confocal using a 60X, 1.4 NA oil immersion objective.

ATP depletion in mammalian cells. Cells were incubated for 30 min in 2 mM sodium azide and 10 mM deoxyglucose in fibronectin-coated glass bottom dishes in HBSS/2% FBS at 37°C and were imaged directly after incubation [52].

Expression in mammalian cells. NPM1-mCherry cell lines were expressed by performing a lentiviral transfection of a cloned SFFV-NPM1-mCherry construct (cloned using In-Fusion HD

Cloning Kit from Clontech to insert NPM1 into a SFFV-mCherry vector). FIB1::GFP expression was performed using FuGENE to transiently transfect 3T3 cells with a CMV-eGFP-FIB1 construct that was a gift from Sui Huang (Addgene plasmid #26673) [53]. RPA194::GFP expression was also performed by transiently transfecting 3T3 cells with a CMV-eGFP-RPA194 construct (a gift from Tom Misteli, Addgene plasmid #17660) [54].

***C. elegans* strain maintenance and imaging.** *C. elegans* lines were maintained at 20°C on NGM plates seeded with OP50 bacteria. Adult hermaphrodites were then anesthetized with 1% levamisole hydrochloride in M9 and imaged on M9-agarose pads using a spinning disk confocal with a 100X/NA 1.4 oil immersion objective.

Generating NPM1::mCherry and FIB1::GFP cross in *C. elegans.* The FIB1::GFP fosmid line was kindly provided by Tony Hyman (MPI-CBG). Crosses were generated by mating FIB1::GFP males with NPM1::mCherry hermaphrodites.

Fluorescence recovery after photobleaching (FRAP) of *X. laevis* nucleoli. Nuclei were injected with mRNA for each nucleolar component, dissected the next day in mineral oil, and placed in an imaging chamber as described above. To deplete ATP, nuclei were injected with 2 mg/mL Apyrase (Sigma) 1-2 hours before dissection. Each nucleolar component was photobleached with a spot 1 μm in diameter and the recovery of fluorescence intensity within the region of interest was obtained for each experiment. Intensity recovery curves were normalized and corrected for photobeaching [55]. To determine the relaxation timescale, τ_f, the recovery curves were fit to the following expression: $I = a - b \cdot e^{-t/\tau_f}$, where a and b are also fit parameters.

FRAP of nucleolar proteins *in vitro.* *In vitro* droplets were photo-bleached with a spot 1 μ m in diameter and the recovery of fluorescence intensity within the region of interest was obtained for

each experiment. Intensity traces were corrected for photo-bleaching, normalized, and fit to an exponential function as above. For aging experiments, FRAP was performed on *in vitro* droplets at different time points as indicated.

FRAP of mammalian cells. 3T3 cells expressing NPM1::mCherry and FIB1::GFP were photo-bleached with a spot ~1μm in diameter and the recovery of fluorescence intensity within the region of interest was obtained for each experiment. Intensity traces were corrected for photo-bleaching, normalized, and fit to the exponential function above. For ATP depletion experiments, cells expressing NPM1::mCherry and FIB1::GFP were incubated in 2 mM sodium azide/10 mM deoxyglucose in HBSS/2% FBS for 30 minutes previous to imaging. Cells were imaged using a heating stage at 37°C.

FRAP of *C. elegans.* *C. elegans* line expressing intestinal FIB1::GFP were maintained at 20 °C20C on NGM plates seeded with OP50 bacteria. L2-L3 larvae or adults were anesthetized with 1% levamisole hydrochloride in M9, placed on M9-agarose pads, and FRAP experiments were performed.

Preparation of non-biological multiphase droplets. Three immiscible liquids were used: DI water, Crisco oil, and silicone oil (viscosity 1,000 cSt, Sigma). For visualization, biotin-4-fluorescein (Biotium) was added to water at 0.1 mg/ml, and Oil Red O (Sigma) was added to Crisco oil at 1 mg/ml. Solutions were made to have a ratio of 5:1:1 of water:silicone oil: Crisco oil. To create multiphase droplets, solutions were vigorously vortexed and pipetted into imaging chambers containing silicone wells (Grace BioLabs).

Microscopy. Experiments for coarsening, surface tension, and *in vivo* characterization of nucleoli of *X. laevis*; *in vitro* experiments with nucleolar proteins, and experiments with *C. elegans* embryos were performed on an inverted Zeiss spinning-disc confocal microscope with

Slidebook software as previously described [22]. Images of Lifeact::GFP network and FRAP experiments were performed on an inverted Nikon laser scanning confocal microscope with a 60X oil immersion objective.

Image Analysis. Custom built software was created in Matlab to perform quantitative image analysis. Code was also adapted from Matlab Multiple Particle Tracking Code (see http://physics.georgetown.edu/matlab/index.html) [56] to apply band pass filters, link nucleoli in three dimensions from volume stacks, and/or track nucleoli in time. ImageJ was used to pseudocolor all images, apply filters, and prepare maximum intensity z-projections.

Design and Implementation of Coarse-Grained Simulations. The lattice-based computer simulations were performed using a coarse-grained description for each molecule and an interaction matrix that defines the effective strengths of inter-module interactions. In keeping with their modular architectures, we modeled FIB1 and rRNA as linear polymers of interaction modules. The pentameric NPM1 was modeled as a branched polymer with five arms. Each arm has three interaction modules anchored to a pre-pentamerized OD (Fig. 5.11A). Each interaction module is represented as a bead (see Fig. 5.11A) and occupies a lattice site such that no two beads can occupy the same site at the same time. The connected architecture is enforced through a linker between beads with a 3D-infinite square well potential. For FIB1 and rRNA, a square well distance of 4 lattice sites was used, and for NPM1, a distance of 2 lattice sites was used. The total number of modules divided by the total number of lattice sites specifies the concentration of modules on the lattice, which has 115 sites to a side. The ternary system comprising of 900 of each of the three polymers starts out in the dispersed phase and is evolved by a collection of $5x10^{10}$ Monte Carlo moves. In the simulations, a bond can form between pairs of modules that occupy adjacent lattice sites. Parameters of the interaction matrix specify if a bond will form

between a pair of interaction modules. This matrix also specifies the effective free energy to be assigned to a given bond. For a pair of modules, the parameters of the interaction matrix are governed by the overall competition between a) the interactions of each module in the pair with the solvent and b) the interactions involving the pair of interest and all other modules in the system. If these two classes of interactions are equivalent in free energy, then no bond will form. If there is an effective preference for the interaction between a pair of modules, then a favorable, negative free energy is assigned to the bond. The configurations of molecules and their positions and orientations with respect to one another were evolved using a Monte Carlo sampling strategy that combines a set of moves including the making and breaking of bonds between modules, pivot moves, crankshaft motions, reptations, and cluster moves of molecules. Pivot moves relocate an end module to a position within its linker length. Crankshaft motions relocate a central module to a position within both its linker lengths. Reptations advance all modules forward like a snake. Cluster moves translate all proteins that are bound together through interactions. Moves that lead to more than one module per lattice site are rejected. The standard Metropolis criterion was used to accept or reject the new configurations that result from bond breaking / making moves. The acceptance criterion for pivot and crankshaft moves was of the form: $\min\{1, N_p N_c^{-1} \exp(-\Delta E)\}$, where N_p and N_c are the number of possible interacting partners, given one to one binding, in the proposed (p) and current (c) positions respectively, and ΔE is the change in energy associated with the proposed move. For reptation moves the acceptance criterion is $\min\{1, (N_p V_p)(N_c/V_c)^{-1} \exp(-\Delta E)\}$, where N_p and N_c are again the number of possible interacting states in the proposed and current states respectively and V_p and V_c are the total number of conformations the module could be placed in the proposed state and current state, respectively. These modifications to classical Metropolis Monte Carlo acceptance ensure the

preservation of microscopic reversibility. Cluster moves do not make or break interactions, nor do they change the internal structure within clusters. Instead, they displace clusters with respect to one another. Cluster moves are always accepted if the move does not engender steric overlap.

5.12 Acknowledgments

We thank members of the Brangwynne laboratory for discussions and Adrienne Fung for preliminary work on surface tension measurements. We acknowledge funding from the Princeton Center for Complex Materials, a MRSEC supported by NSF Grant DMR 1420541. This work was also supported by an NIH Director's New Innovator Award (1DP2GM105437-01) (C.P.B.), an NSF CAREER award (1253035) (C.P.B.), a Helen Hay Whitney Fellowship (N.V.), an NSF grant (MCB 1121867) (R.V.P.), an NIH grant (5RO1NS056114) (R.V.P.), an NIH grant (5R01GM115634) (R.W.K.), an NCI Cancer Center Support grant (P30CA21765 at St. Jude Children's Research Hospital) (R.W.K.), and ALSAC (R.W.K.).

5.13 References

1. Feric, M., et al., *Coexisting Liquid Phases Underlie Nucleolar Subcompartments.* Cell, 2016. **165**(7): p. 1686-97.
2. Balagopal, V. and R. Parker, *Polysomes, P bodies and stress granules: states and fates of eukaryotic mRNAs.* Current opinion in cell biology, 2009. **21**(3): p. 403-408.
3. Spector, D.L., *Nuclear domains.* Journal of cell science, 2001. **114**(16): p. 2891-2893.
4. Mitrea, D.M., et al., *Nucleophosmin integrates within the nucleolus via multi-modal interactions with proteins displaying R rich linear motifs and rRNA.* Elife, 2016: p. e13571.

5. Brangwynne, C.P., et al., *Germline P granules are liquid droplets that localize by controlled dissolution/condensation.* Science, 2009. **324**(5935): p. 1729-1732.

6. Nott, T.J., et al., *Phase transition of a disordered nuage protein generates environmentally responsive membraneless organelles.* Molecular cell, 2015. **57**(5): p. 936-947.

7. Li, P., et al., *Phase transitions in the assembly of multivalent signalling proteins.* Nature, 2012. **483**(7389): p. 336-340.

8. Weber, S.C. and C.P. Brangwynne, *Inverse size scaling of the nucleolus by a concentration-dependent phase transition.* Current Biology, 2015. **25**(5): p. 641-646.

9. Berry, J., et al., *RNA transcription modulates phase transition-driven nuclear body assembly.* Proceedings of the National Academy of Sciences, 2015. **112**(38): p. E5237-E5245.

10. Wippich, F., et al., *Dual specificity kinase DYRK3 couples stress granule condensation/dissolution to mTORC1 signaling.* Cell, 2013. **152**(4): p. 791-805.

11. Molliex, A., et al., *Phase separation by low complexity domains promotes stress granule assembly and drives pathological fibrillization.* Cell, 2015. **163**(1): p. 123-133.

12. Patel, A., et al., *A liquid-to-solid phase transition of the ALS protein FUS accelerated by disease mutation.* Cell, 2015. **162**(5): p. 1066-1077.

13. Elbaum-Garfinkle, S., et al., *The disordered P granule protein LAF-1 drives phase separation into droplets with tunable viscosity and dynamics.* Proceedings of the National Academy of Sciences, 2015: p. 201504822.

14. Wang, J.T., et al., *Regulation of RNA granule dynamics by phosphorylation of serine-rich, intrinsically disordered proteins in C. elegans.* Elife, 2015. **3**: p. e04591.

15. Hennig, S., et al., *Prion-like domains in RNA binding proteins are essential for building subnuclear paraspeckles.* The Journal of cell biology, 2015. **210**(4): p. 529-539.

16. Derenzini, M., et al., *Nucleolar size indicates the rapidity of cell proliferation in cancer tissues.* The Journal of pathology, 2000. **191**(2): p. 181-186.

17. Frank, D.J. and M.B. Roth, *ncl-1 is required for the regulation of cell size and ribosomal RNA synthesis in Caenorhabditis elegans.* The Journal of cell biology, 1998. **140**(6): p. 1321-1329.

18. Boisvert, F.-M., et al., *The multifunctional nucleolus.* Nat Rev Mol Cell Bio, 2007. **8**(7): p. 574-585.

19. Jain, S., et al., *ATPase-Modulated Stress Granules Contain a Diverse Proteome and Substructure.* Cell, 2016.

20. Hubstenberger, A., et al., *Translation repressors, an RNA helicase, and developmental cues control RNP phase transitions during early development.* Developmental cell, 2013. **27**(2): p. 161-173.

21. Brangwynne, C.P., T.J. Mitchison, and A.A. Hyman, *Active liquid-like behavior of nucleoli determines their size and shape in Xenopus laevis oocytes.* Proc Natl Acad Sci U S A, 2011. **108**(11): p. 4334-4339.

22. Feric, M. and C.P. Brangwynne, *A nuclear F-actin scaffold stabilizes ribonucleoprotein droplets against gravity in large cells.* Nature Cell Biology, 2013. **15**(10): p. 1253-1259.

23. Feric, M., C.P. Broedersz, and C.P. Brangwynne, *Soft viscoelastic properties of nuclear actin age oocytes due to gravitational creep.* Scientific reports, 2015. **5**.

24. Than, P., et al., *Measurement of interfacial tension between immiscible liquids with the spinning road tensiometer.* Journal of colloid and interface science, 1988. **124**(2): p. 552-559.

25. Aarts, D.G., M. Schmidt, and H.N. Lekkerkerker, *Direct visual observation of thermal capillary waves.* Science, 2004. **304**(5672): p. 847-850.

26. Zhang, H., et al., *RNA controls PolyQ protein phase transitions.* Molecular cell, 2015. **60**(2): p. 220-230.

27. Parry, B.R., et al., *The bacterial cytoplasm has glass-like properties and is fluidized by metabolic activity.* Cell, 2014. **156**(1): p. 183-194.

28. Kroschwald, S., et al., *Promiscuous interactions and protein disaggregases determine the material state of stress-inducible RNP granules.* Elife, 2015. **4**: p. e06807.

29. Mitrea, D.M., et al., *Structural polymorphism in the N-terminal oligomerization domain of NPM1.* Proceedings of the National Academy of Sciences, 2014. **111**(12): p. 4466-4471.

30. Rubinstein, M. and R.H. Colby, *Polymer physics*. NEW YORK: Oxford University. 2003.

31. Neeson, M.J., et al., *Compound sessile drops.* Soft Matter, 2012. **8**(43): p. 11042-11050.

32. Pederson, T., *The nucleolus.* Cold Spring Harbor perspectives in biology, 2011. **3**(3): p. a000638.

33. Thiry, M. and D.L. Lafontaine, *Birth of a nucleolus: the evolution of nucleolar compartments.* Trends in cell biology, 2005. **15**(4): p. 194-199.

34. Tschochner, H. and E. Hurt, *Pre-ribosomes on the road from the nucleolus to the cytoplasm.* Trends in cell biology, 2003. **13**(5): p. 255-263.

35. Henras, A., et al., *The post-transcriptional steps of eukaryotic ribosome biogenesis.* Cellular and Molecular Life Sciences, 2008. **65**(15): p. 2334-2359.

36. Gigova, A., et al., *A cluster of methylations in the domain IV of 25S rRNA is required for ribosome stability.* RNA, 2014. **20**(10): p. 1632-1644.

37. Kruger, T., H. Zentgraf, and U. Scheer, *Intranucleolar sites of ribosome biogenesis defined by the localization of early binding ribosomal proteins.* The Journal of cell biology, 2007. **177**(4): p. 573-578.

38. Weber, S.C. and C.P. Brangwynne, *Getting RNA and protein in phase.* Cell, 2012. **149**(6): p. 1188-1191.

39. Lin, Y., et al., *Formation and maturation of phase-separated liquid droplets by RNA-binding proteins.* Molecular cell, 2015. **60**(2): p. 208-219.

40. Xiang, S., et al., *The LC Domain of hnRNPA2 Adopts Similar Conformations in Hydrogel Polymers, Liquid-like Droplets, and Nuclei.* Cell, 2015. **163**(4): p. 829-839.

41. Ochs, R., et al., *Fibrillarin: a new protein of the nucleolus identified by autoimmune sera.* Biology of the Cell, 1985. **54**(2): p. 123-133.

42. Foty, R.A., et al., *Surface tensions of embryonic tissues predict their mutual envelopment behavior.* Development, 1996. **122**(5): p. 1611-1620.

43. Gall, J.G., *Cajal bodies: the first 100 years.* Annual review of cell and developmental biology, 2000. **16**(1): p. 273-300.

44. Shav-Tal, Y., et al., *Dynamic sorting of nuclear components into distinct nucleolar caps during transcriptional inhibition.* Molecular biology of the cell, 2005. **16**(5): p. 2395-2413.

45. Wachtler, F. and A. Stahl, *The nucleolus: a structural and functional interpretation.* Micron, 1993. **24**(5): p. 473-505.

46. Kedersha, N., et al., *Stress granules and processing bodies are dynamically linked sites of mRNP remodeling.* The Journal of cell biology, 2005. **169**(6): p. 871-884.

47. Buchan, J.R. and R. Parker, *Eukaryotic stress granules: the ins and outs of translation.* Molecular cell, 2009. **36**(6): p. 932-941.

48. Dumont, J.N., *Oogenesis in Xenopus laevis (Daudin). I. Stages of oocyte development in laboratory maintained animals.* J Morphol, 1972. **136**(2): p. 153-79.

49. Zacharias, D.A., et al., *Partitioning of lipid-modified monomeric GFPs into membrane microdomains of live cells.* Science, 2002. **296**(5569): p. 913-916.

50. Gall, J.G. and Z. Wu, *Examining the contents of isolated Xenopus germinal vesicles.* Methods, 2010. **51**(1): p. 45-51.

51. Hansen, F., *Surface tension by image analysis: Fast and automatic measurements of pendant and sessile drops and bubbles.* Journal of colloid and interface science, 1993. **160**(1): p. 209-217.

52. Brangwynne, C.P., F. MacKintosh, and D.A. Weitz, *Force fluctuations and polymerization dynamics of intracellular microtubules.* Proceedings of the National Academy of Sciences, 2007. **104**(41): p. 16128-16133.

53. Chen, D. and S. Huang, *Nucleolar components involved in ribosome biogenesis cycle between the nucleolus and nucleoplasm in interphase cells.* The Journal of cell biology, 2001. **153**(1): p. 169-176.

54. Dundr, M., et al., *A kinetic framework for a mammalian RNA polymerase in vivo.* Science, 2002. **298**(5598): p. 1623-1626.

55. Phair, R.D., S.A. Gorski, and T. Misteli, *Measurement of dynamic protein binding to chromatin in vivo, using photobleaching microscopy.* Methods in enzymology, 2003. **375**: p. 393-414.

56. Crocker, J.C. and D.G. Grier, *Methods of digital video microscopy for colloidal studies.* J Colloid Interface Sci, 1996. **179**(1): p. 298-310.

Chapter 6

Order and Disorder in Protein Biomaterial Design

This chapter is adapted from an article under preparation. Stefan Roberts, Jeffery Schaal, Kan (Jonathan) Li, Kai Wang, Andrew Hunt, Vincent Miao, Terrence Oas, and Ashutosh Chilkoti designed and conducted the experiments. Tyler S. Harmon and Rohit V. Pappu developed the coarse-grained framework. Tyler S. Harmon performed and analyzed the simulations.

6.1 Introduction

Both purely crystalline and amorphous materials have been extensively studied for their interesting properties, but they comprise a very small portion of the total materials space. Most material properties are a consequence of the interplay between their ordered and disordered domains. This phenomenon is one of the hallmarks of biological materials —for example silk fibers owe their extraordinary attributes to the interactions of ordered and disordered domains at the inter- and intra- molecular level[1]. With the recent expansion of research on intrinsically disordered proteins (IDPs), the importance of disorder-order interactions has become further undeniable[2, 3]. To understand how this interplay creates macroscopic material properties, ordered and disordered nanoscale modules have to be synthesized with molecular precision. The emergence of genetically encoded synthesis of peptide polymers finally makes it possible to design building blocks with this level of control over sequence and structure[4, 5]. Motivated by the advancement of molecular tools and the increasing applications of protein materials, we have

rationally designed modular, protein biopolymers in which we precisely tune their internal order and disorder to elucidate new rules for materials design.

6.2 Polymer Library Design

Elastin-like polypeptides (ELPs) are a family of recombinant proteins that have received significant interest in the past decade. They are based on a consensus sequence from the disordered regions of the IDP tropoelastin and exhibit a tunable lower critical solution temperature (LCST) phase behavior[5-8] that has been used for a number of applications including protein purification[9], drug delivery[10], and tissue engineering[11]. ELPs have been characterized as models of elastomeric disorder and their intrinsic disorder is primarily responsible for their sharp LCST behavior[8, 12]. Polyalanine domains, on the other hand, offer the highest degree of helix structural stability barring the inclusion of stabilizing side chain interactions[13, 14]. They have been extensively studied due to their biological prevalence – the third most abundant homopeptide repeat in eukaryotes – and their propensity towards aggregation[14-16]. Polyalanine helices are particularly important in tropoelastin where they combine with disordered domains to produce the incredible material properties that make elastin such a valuable component of the extracellular matrix[17-20]. We hypothesized that recombinant polymers composed of alternating ELP and polyalanine domains, which mimics the exon composition and organization of tropoelastin[21-25], would similarly produce biomaterials with unique, tunable properties. As an ideal elastomeric disordered protein and an ideal helix, ELP and polyalanine provide contrasting extremes contained within a design simplicity that allows us to parse the effect of both elements while retaining broad applicability to other sequences.

Four polyalanine helices (H1,2,3,5) with different charge distributions were incorporated into three different ELPs (E1-3) of varying side chain hydrophobicities at either 7.25%, 12.5%,

25%, or 50% of the total amino acid number (Fig. 6.1a). The compositions of the polyalanine domains were chosen to maximize helicity while controlling the hydrophilicity of the peptide domains through charge-charge interactions. ELP domains with alanine and valine guest residues were chosen to span a range of LCSTs at temperature suitable for *in vivo* injection. Increasing the alanine content increases the hydrophilicity and therefore the LCST of the polymers. The naming convention for our partially ordered polymers (POPs) used throughout this document specifies the ELP sequence (E*X*), the helix sequence (H*Y*), and percent helical content (*Z* %): E*X*-H*Y*-*Z*%. For example, a 400 amino acid polymer (~33kDa) encoding two GA_{25} helices into VPGVG is referred to herein as E1-H1-12.5%.

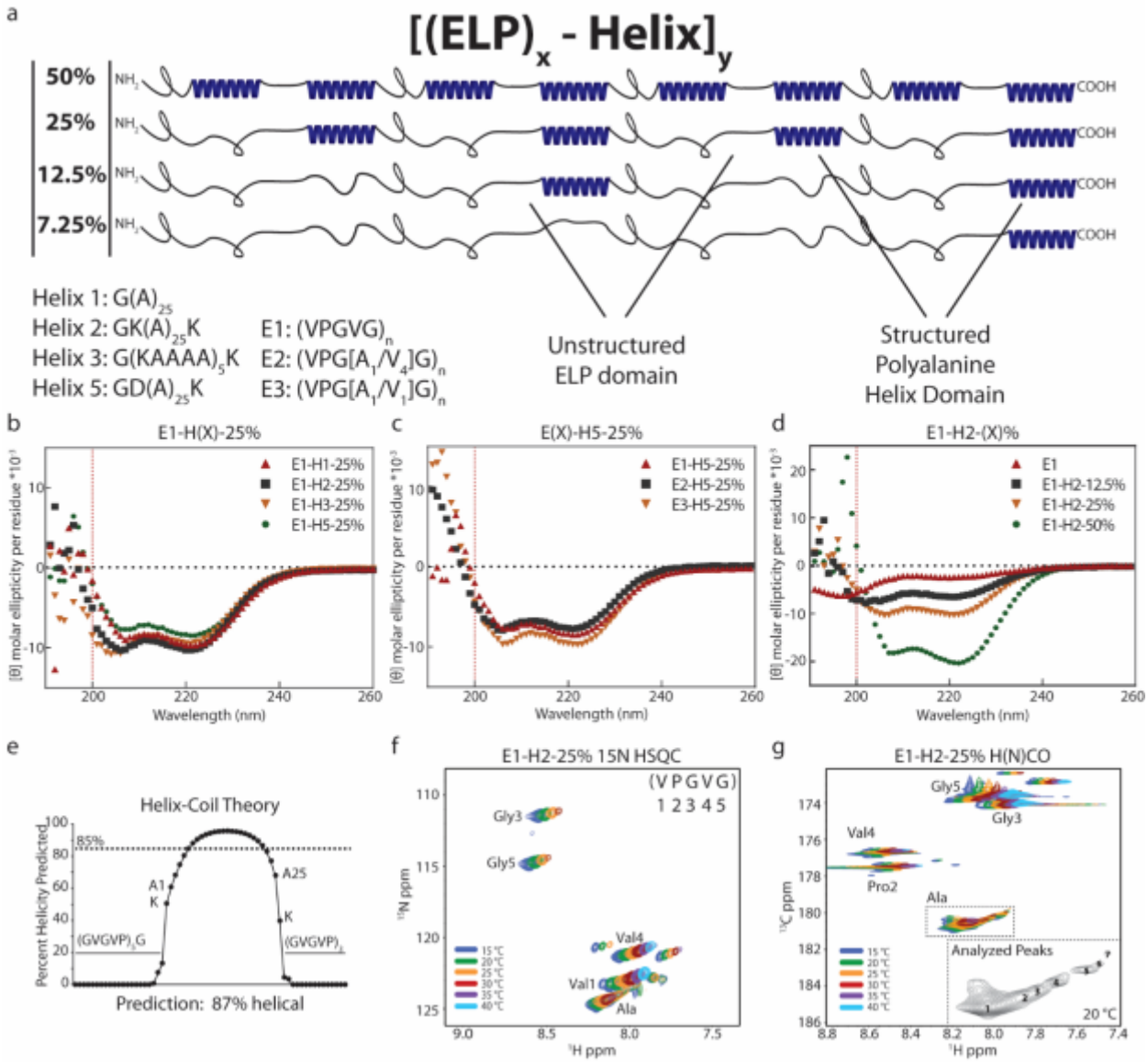


Figure 6.1: Partially Ordered Polymer Library and Structural Characterization. (a) Recombinant POPs were constructed with 3 ELP components and 4 polyalanine helices at amino acid percentages up to 50%. CD reveals definitive helical peaks at 222 and 208 nm, with peak amplitudes minimally altered by (b) polyalanine domain and (c) ELP but highly dependent on (d) total alanine content (dynode voltage >500 left of red line). (e)This structural signature is consistent with helix-coil predictions (Agadir). (f) ^{15}N-HSQC and (g) H(N)CO (residue labels are the associated C' of the previous residue) 2D solution NMR spectra for E1-H2-25% were used to more precisely quantify total structural content. Each polyalanine domain was determined to have an average helicity of 90% (Sup. Methods).

6.3 Structural Characterization

We hypothesized that once encoded into an ELP, polyalanine domains would retain their high degree of helicity, and used circular dichroism (CD) to determine the structural features of

each polymer. All POPs show the negative ellipticity peaks at 222nm and 208nm (Fig. 6.1b-d) characteristic of α-helices. The magnitudes of these peak positions is largely independent of polyalanine and ELP domain composition but highly dependent on the total polyalanine percentage of each polymer. The helices are thermally stable and show minimal melting at higher temperatures. Because quantitative analysis of CD data for disordered proteins is known to be inaccurate[27], we also performed a series of 2D-solution NMR experiments to more precisely determine the percentage of helicity. Though the repetitive and proline rich nature of ELPs increases resonance assignment complexity, assignment for key amino acids was still feasible using combinations of triple resonance NMR spectra (Fig. 6.1f-g). The backbone carbonyl carbon chemical shifts of the alanine peaks in the H(N)CO spectrum – a particularly sensitive spectrum to detect secondary structure changes – were used to quantify helicity. Based on these chemical shifts, 90% of the alanines within each helical domain (H2) are predicted to be in a helical conformation at 20 °C. This result is supported by helix-coil transition theory prediction algorithms (Fig. 6.1e)[28-31], and the temperature dependent change of the chemical shifts of backbone carbonyl carbon. Given the similarity in CD structural signatures (Fig 6.1b), the remaining helical compositions can be confidently approximated to a similar degree of structure.

6.4 Sharp Phase Behavior and Tunable Hysteresis

ELPs possess reversible LCST behavior, the ability to cycle between soluble and aggregated states with no permanent changes, where the transition of temperature of heating (T_t-heating) is identical to the transition temperature upon cooling (T_t-cooling). As the disordered nature of ELPs is necessary for their sharp phase behavior, we anticipated that the incorporation of highly ordered domains could definably alter this behavior. We assessed the thermal phase

transition of our POPs by monitoring their optical translucence during steady heating. Remarkably, all proteins demonstrate very sharp (<1-2°C) phase transitions at precise temperatures, even when composed of 50% total helicity (Fig. 6.2a-d). These transition temperatures vary depending on the specific ELP and helix composition due to differences in hydrophilicity and charge, but all polymers possess the sharp phase behavior characteristic of fully disordered ELPs. When the POPs were subsequently cooled, the POPs were likewise found to reversibly disassemble.

One aspect of the POPs thermal behavior, however, was of particular interest – the marked downshift in the dissolution temperature (T_t-Cooling) from the original T_t-heating. This thermal hysteresis, defined as the difference between T_t-heating and T_t-cooling (ΔT_t), has been observed in other recombinant polymers[32-37]. The increased stability afforded by this property has been advantageous for development of hyper-stable nano/micro-particles for controlled drug release and for mechanically stabilizing protein scaffolds for tissue engineering[34, 36]. However, the limited number of sequences and limited control over the hysteretic range of those sequences has severely limited their potential. In contrast, the hysteresis for our POPs can be precisely controlled as it is directly correlated with total helical content (Fig. 6.2a-b) and inversely correlated with amount of charge on the helix side chains (Fig. 6.2c-e). Importantly, once fully solvated, polymers return to their original state and can be cyclically heated and cooled with no permanent alterations (Fig. 6.3a-b). By incorporating a helix with sufficient charge repulsion, such as H3, hysteresis can be even be eliminated altogether. Hysteresis is also independent of both heating and cooling rates, and polymers heated and then cooled to the hysteretic range below their T_t-heating show no change in solvation after 24hrs (Fig. 6.3c-e). Subsequent cooling below the T_t-cooling after 24hrs causes rapid dissolution.

Traditional ELP transition temperatures scale as a function of the logarithm of polymer concentration[5, 7]. In our POP system, T_t-heating was found to scale logarithmically with polymer concentration, in accordance with traditional ELP behavior. However, T_t-cooling was found to be concentration independent (Fig. 6.2d-f). Altering the guest residues of the ELP backbone, and therefore the overall hydrophobicity of the polymer, adjusts the T_t-heating appropriately, but does not change the T_t-cooling (Fig. 6.2f). These observations suggest that helix composition is the primary determinant for dissolution upon cooling. Our design provides a system whereby the sequences that drive initial aggregation are distinct from the sequences controlling dissolution. This unique separation allows orthogonal tuning of heating and cooling transition temperatures.

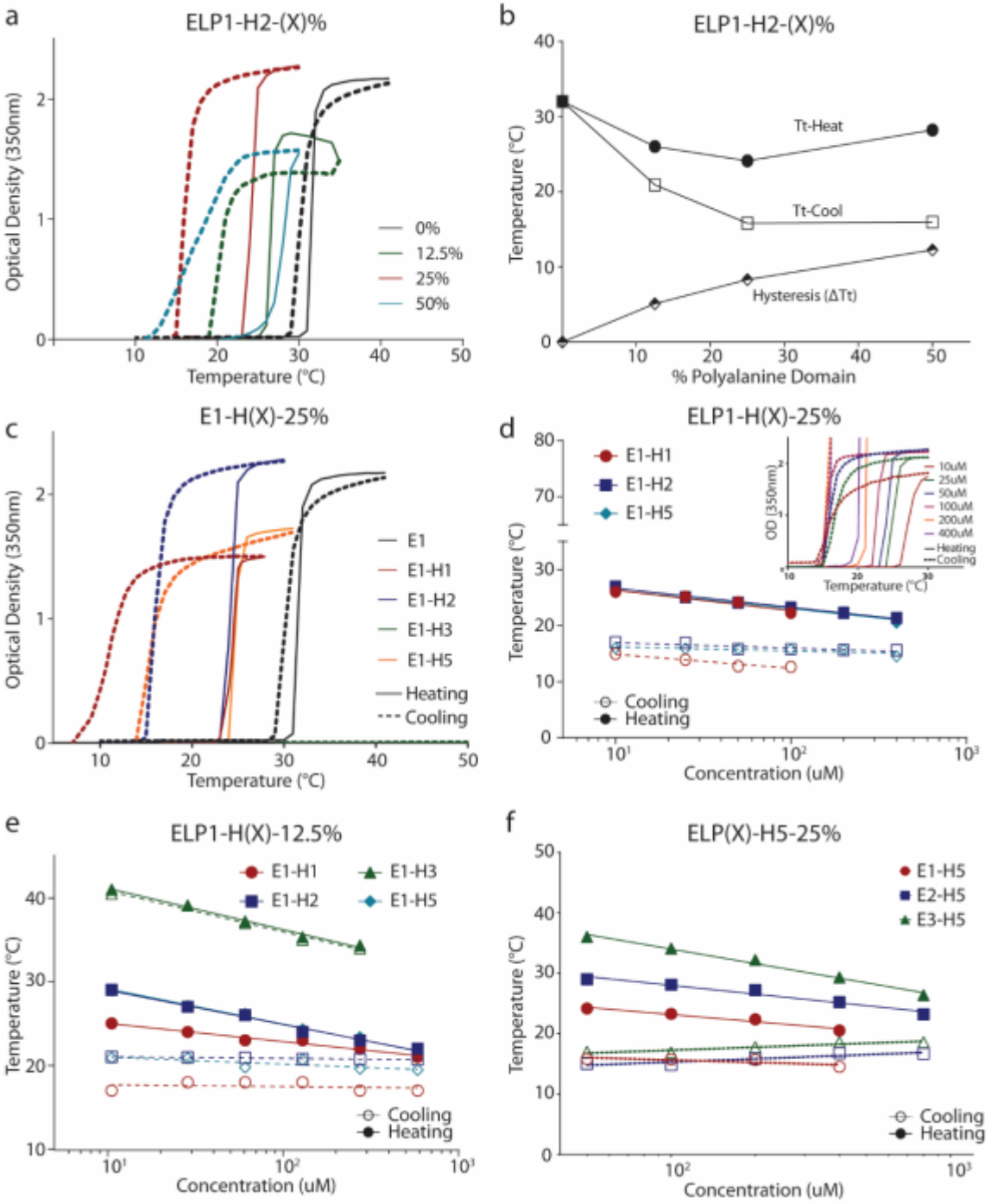


Figure 6.2: Turbidity and Hysteresis. (a) Optical density measurements as a function of temperature show sharp, reversible phase behavior and a difference in the Tt-heating and Tt-cooling (hysteresis) which (b) scales as a function of included polyalanine domains. (c-e) Hysteresis is also dependent on the composition (charge distribution) of the polyalanine domains with an increase in charge producing a decrease in hysteresis. (d-f) The Tt-cooling is dependent on helix composition and independent of concentration though (f) the Tt-heating can be independently tuned by altering the ELP composition. Optical density measurements at 350nm in PBS at 50uM unless otherwise indicated. Heating and cooling rates were kept at 1°C/min. OD amplitudes are non-interpretable due to difference in aggregate formation and settling.

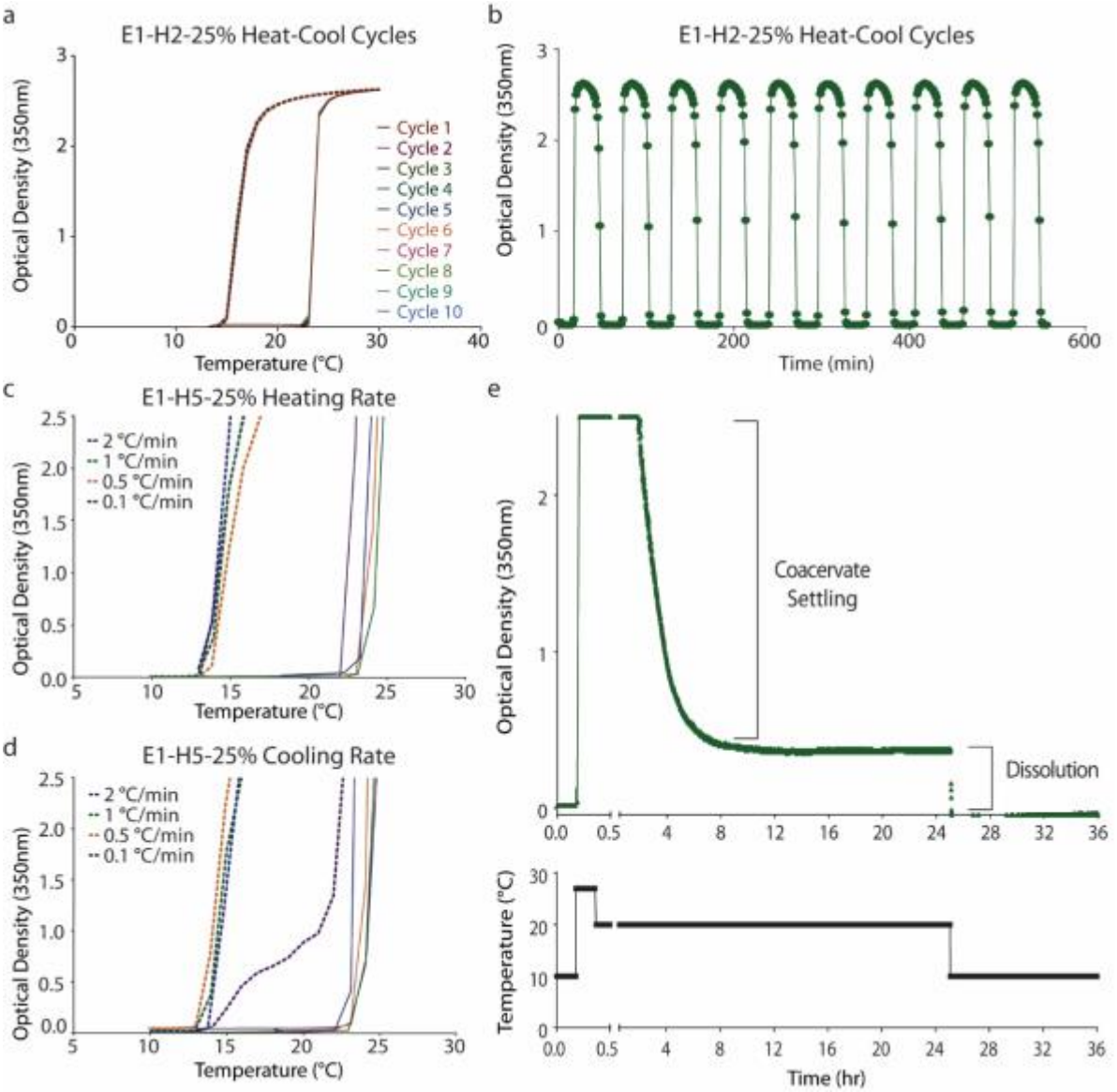


Figure 6.3: Kinetics of Hysteresis. (a-b) Despite their hysteretic nature, polymers are capable of fully recovering from heating and cooling cycles. Ten cycles show no change in transition temperatures. Altering the (c) heating rates and (d) cooling rates also does not change the phase behavior, though some settling occurs at slower cooling rates. (e) E1-H5-25% (50uM, PBS) shows no recovery for 24 hrs when heated and cooled to the hysteretic range above the Tt-cooling. Subsequent cooling after 24 hours shows rapid dissolution.

6.5 A Model for the Mechanism of Hysteresis

The mechanistic underpinnings of thermal hysteresis have commonly been attributed to changes in the secondary structures adapted by polymers[37, 38]. Polyalanine is known to adopt coil, helical, and beta configurations[39], so we first analyzed our system to determine if a

change in secondary structure upon aggregation was driving this behavior. Temperature controlled CD spectra of a hydrophilic POP (E1-H3-25%) indicate that, in the absence of aggregation, the polymers will retain a high degree of helicity even up to 65 ˚C (Fig. 6.4a). Those polymers that do aggregate show CD spectral distortions (Fig. 6.4b-c) consistent with those observed for aggregates of other helical peptides and of tropoelastin[40-43]. These spectral shifts strongly suggest the presence of helices within the protein aggregates. We also evaluated the phase behavior of our polymers in urea (up to 8M), expecting its introduction to minimize hysteresis if a shift in secondary structure was responsible. While urea did predictably increase the T_t-heating[44], it did not have a significant effect on the dissolution temperatures (Fig. 6.5) controlled by the helical domains. These results suggest that the helical rigidity itself is not the driving force for hysteresis; rather, the structural components act as a presentation platform for additional interactions among side chains along the helix. This mechanism is consistent with coacervation mechanics of tropoelastin, in which polyalanine domains increase in helicity to stabilize side-chain interactions for crosslinking[45, 46].

Given the intrinsic tendency of alanine domains to aggregate[15, 16] and the persistence of helices within the protein aggregates, we propose that hysteresis is driven by helical clustering and have further explored the interaction mechanisms using coarse grain molecular dynamics simulations. We used a phenomenological model separating the protein domains into two categories of five amino acid “beads”: polyalanine (AAAAA) and ELP (VPGVG). The interactions energies between polyalanine domains (E_{AA}) are always the most preferred (intrinsic alanine aggregation), and the interaction energies between ELP domains (E_{EE}) change with temperature with weakly attractive energies below the LCST and strongly attractive above. Interactions between polyalanine domains and ELPs (E_{EA}) were always considered positive. We

simulated a hysteretic cycle for 50 polymers of 25% helicity in a 25nm radius spherical box. The results (Fig. 6.6) suggest that POPs move through four separate stages during a complete heating and cooling cycle. (1) Below their LCST, POPs are in a state of isolated oligomers in which local alanine domains have clustered, but these clusters remain isolated and sufficiently solvated by their ELP domains. (2) Above the LCST these localized aggregate clusters dock with one another due to the increased favorability of ELP hydrophobic interactions. (3) Given sufficient time, we expect the alanine domains to exchange with neighboring docked clusters such that single POPs can span more than the single cluster with which they have docked. This has the effect of entangling the aggregate clusters into a percolated network. Swapping of alanine domains between clusters is feasible because of the high density in the docked state and thermodynamically favored through the entropy of mixing. Additionally, as the temperature is increased further and the repulsive term of the ELPs continues to decrease, a second reversible transition becomes favored where docked spherical clusters convert into linear aggregates that are denser and should be expected to be less dynamic than the entangled spherical aggregates. (4) Once cooled to below the Tt-heating, the entanglement of the aggregates prevents dissolution of the ELP domains, resulting in an entangled oligomers state. Unlike the fast and irreversible transition from docked aggregates to entangled aggregates (2-3), transitions between entangled oligomers and isolated oligomers are expected to be slow. A sufficient drop in ELP interaction energy (additional cooling) will eventually solvate the POPs, diluting the clusters, and returning them to their original state.

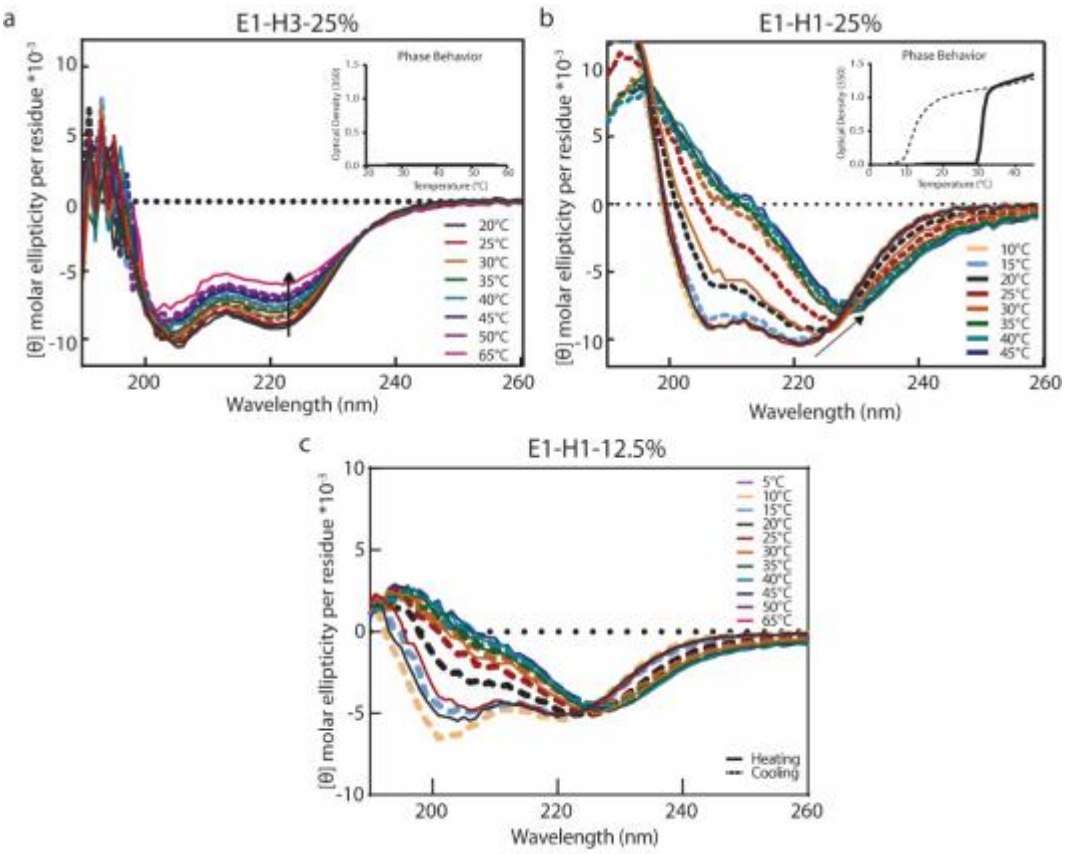


Figure 6.4: Temperature Dependent CD. (a) E1-H3-25%, which does not transition at low temperatures, shows the preservation of helical signature peaks at high temperatures with some loss in peak amplitudes. (b-c) E1-H1-25%, which does transition, shows a spectral shift consistent with distortions for helical peptides at the expected transition temperature. This polymer also shows an isodichroic point at 225nm for both 12.5 and 25%. 10uM, water, 1mm path length for all experiments.

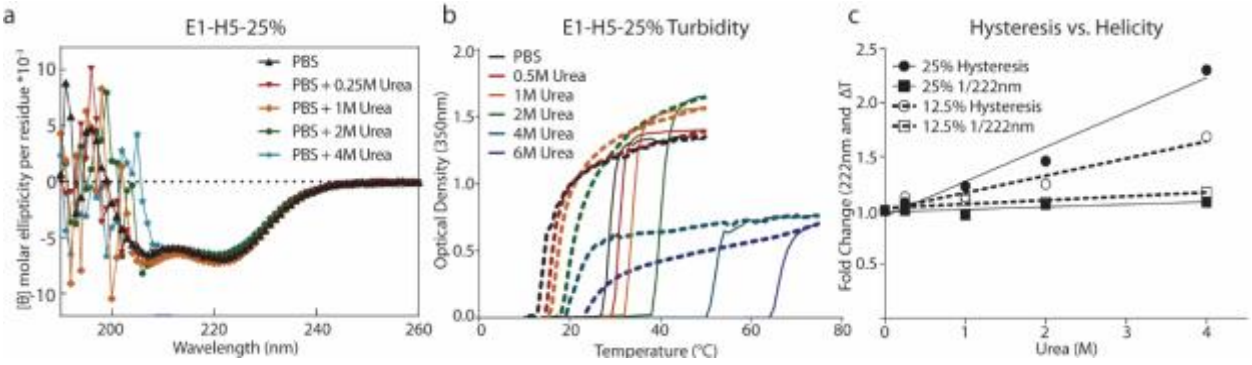


Figure 6.5: Impact of Urea on Helicity and Hysteresis. (a) Up to 4M Urea did not appreciably alter the CD peak amplitudes for E1-H5-25% (10uM). Dynode voltages >500 at wavelengths ranging from 200-215 dependent on Urea concentration. (b) Despite not altering the helicity, Urea dramatically increases the Tt-heating and has a minor effect on the Tt-cooling. (c) Normalized changes in 222nm peaks and hysteresis (ΔT) illustrate that, while helicity and hysteresis are related, they are not always correlated.

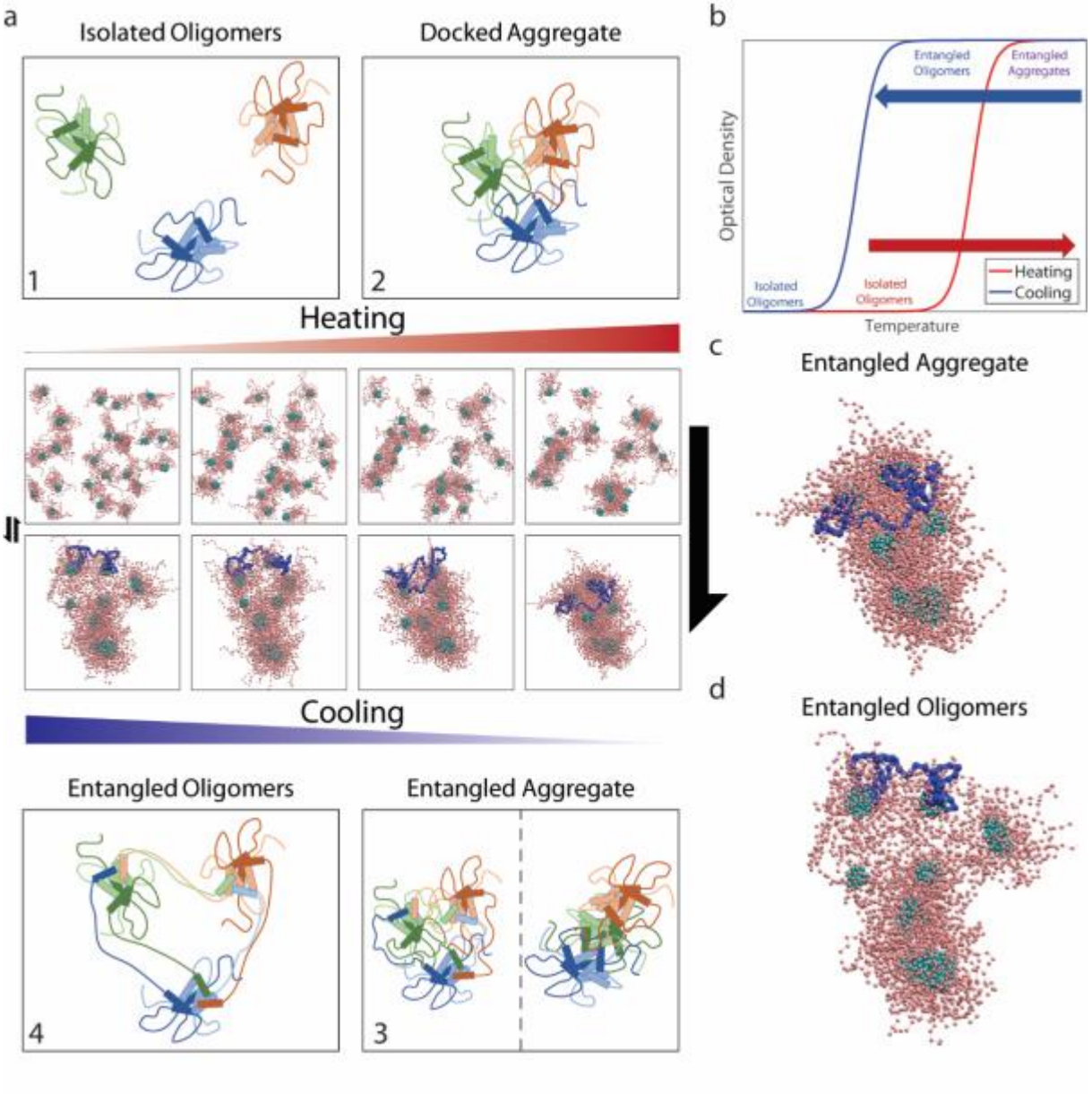


Figure 6.6: Proposed Mechanism for Hysteresis. Simulations of the hysteretic cycle were performed using a coarse grained model. Heating and cooling were achieved by modulating the interaction strengths between ELP domains. (a) Snapshots extracted from a phenomenological simulation of POPs shown in the middle, surrounded by cartoon representations of the four states observed for POP during heating and cooling. Rod-like objects represent alanine domains and string-like tethers represent ELPs. The colors indicate their initial cluster with shading indicating different proteins in the same initial cluster. The one-sided arrows provide a pictorial summary of the expected rates for transitions between different states (fast for 2-3 and slow for 4-1). Within entangled aggregates we observe two types of morphologies viz., entangled spheres or entangled cylinders. There is a reversible spheres to cylinders transition at even higher temperatures. (b) A simplified representation of experimental data is annotated by the species populating each regime. The ordinate is labeled as a measure of optical density consistent with experimental work. (c-d) Enlarged snapshots from the cooling arm of panel (a) demonstrate that the highlighted POP is not able to isolate itself into a single cluster and that the decrease in aggregate density is limited by the presence of domain swapped proteins.

6.6 Formation of Solid-Like, Fractal Networks

The macroscale properties of POP aggregates also indicate a mechanism for aggregation distinct from the liquid coacervation associated with disordered ELPs. Rather than a turbid suspension, POPs transition into mechanically stable, opaque aggregates. Frequency sweeps in the linear viscoelastic region of ELPs show the loss modulus (G") (23Pa, 1 Hz, 10mg/ml) to be greater than the storage modulus (G') (8.0 Pa, 1 Hz, 10mg/ml) and both to be proportional to frequency—behavior consistent with liquid-like coacervates (Fig 6.7a-b)[47, 48]. Similar measurements for POPs reveal G' (12.2 kPa, 1 Hz, 10mg/ml) to be much greater than G" (0.36 kPa, 1Hz, 10mg/ml) and independent of frequency—behavior typical of more solid-like materials (Fig 6.7b-c)[47, 48]. At equivalent concentrations, G' for the POPs was up to four orders of magnitude higher than that for ELPs. Oddly, moduli for all polymers at all concentrations converge at high frequencies, though we cannot yet offer an explanation for this phenomenon. Consistent with these observations, POPs display a high viscosity with plastic, shear-thinning-flow, while ELPs behave as a Newtonian fluid (Fig. 6.7d). The shear thinning slope for POPs was unusually high (-0.95) for long-chain polymers, though this outcome is consistent with previously reported values for tropoelastin exon networks[47]. The high influence of shear suggests some rupture due to the absence of covalent crosslinks.

We also observe reversible syneresis with POPs in which water is expelled from the aggregates, resulting in cracking, hardening, and shrinking as the temperature is increased (Fig. 6.8a). ELP coacervates have also been observed to shrink at higher temperatures as solvent quality decreases and interactions between the hydrophobic domains become more preferable; however, this property is not observed in bulk solutions of ELPs since the coacervates remain

largely isolated in a colloidal suspension. Syneresis on an observable scale suggests percolated crosslinking interactions between polymers and is likely a result of the helical clustering supported by our simulations. Of note, the addition of charge repulsion along the backbone of the helix in H3, which eliminates hysteresis, also restores solution-like coacervation behavior.

The incorporation of structural components also has a profound effect on microscale phase separation (Fig. 6.8b-c). Soluble ELPs form submicron-sized aggregates which continually mature to form larger immiscible spheres. In contrast, POPs undergo an arrested phase separation into porous, fractal protein networks. By monitoring fluorescence recovery after photobleaching (FRAP), we determined that the porous networks are robust and kinetically stable with minimal fluorescence recovery observed after 30 minutes (Fig. 6.9a). This high degree of stability is due to the more solidified state of the protein rich phase which minimizes chain mobility within the aggregated network. There is slightly more recovery for 12.5% networks, but the unrecovered fraction remains high (86%). Moreover, the porosity of the network can be controlled by modulating polymer concentration. Using three dimensional reconstructions of confocal microscopy images, we evaluated the effects of concentration on total void volume, defined as the non-protein rich phase of the arrested network (Fig. 6.9b-c). Within a range of 50uM (0.2 mg/ml) to 800uM (2.6 mg/ml) void volume can be tuned to between 90% and 60%. While further increasing the concentration of the proteins appears to produce networks with lower void volume and tighter pores, specific quantification using our methodology was not possible as the feature size approached the maximum resolution of the microscope. Although 12.5% polymers are subject to more thermal fluctuation and increased mobility, the total helical percentage does not significantly factor into network void volume.

Because the phase separation of these particles occurs on a length scale below the diffraction limit and is unreachable with traditional light microscopy, we used structured illumination microscopy (SIM), a superresolution light microscopy technique[49, 50], to better characterize the architecture of the networks in a physiological environment. SIM reveals them to be comprised of mesoscale polymer aggregates no larger than 200nm interconnected with a "pearl-necklace" like architectures (Fig. 6.8d), and this mesoscale architecture is consistently observed across multiple polymer compositions. This observation is suggestive of a two-stage aggregation process. The polymers initially nucleate in a similar fashion to their disordered counterparts (T_t-heating is driven by the disordered portion). Rather than coalesce, however, aggregates rapidly link, forming the observed fractal networks. Our MD simulations also predicted a two-stage process on the nanoscale (aggregate docking and subsequent entanglement), and it is reasonable to theorize that similar entanglements also occur on a meso to micro-scale. This type of aggregation is also mirrored in tropoelastin which has been well documented to undergo a multistage aggregation process[46, 51]. This process includes an initial hydrophobic coacervation into spherical droplets ranging from 200nm (each containing 10^4 proteins) to 6um and subsequent maturation into porous networks or fibers due to interactions between crosslinking domains[46, 51, 52].

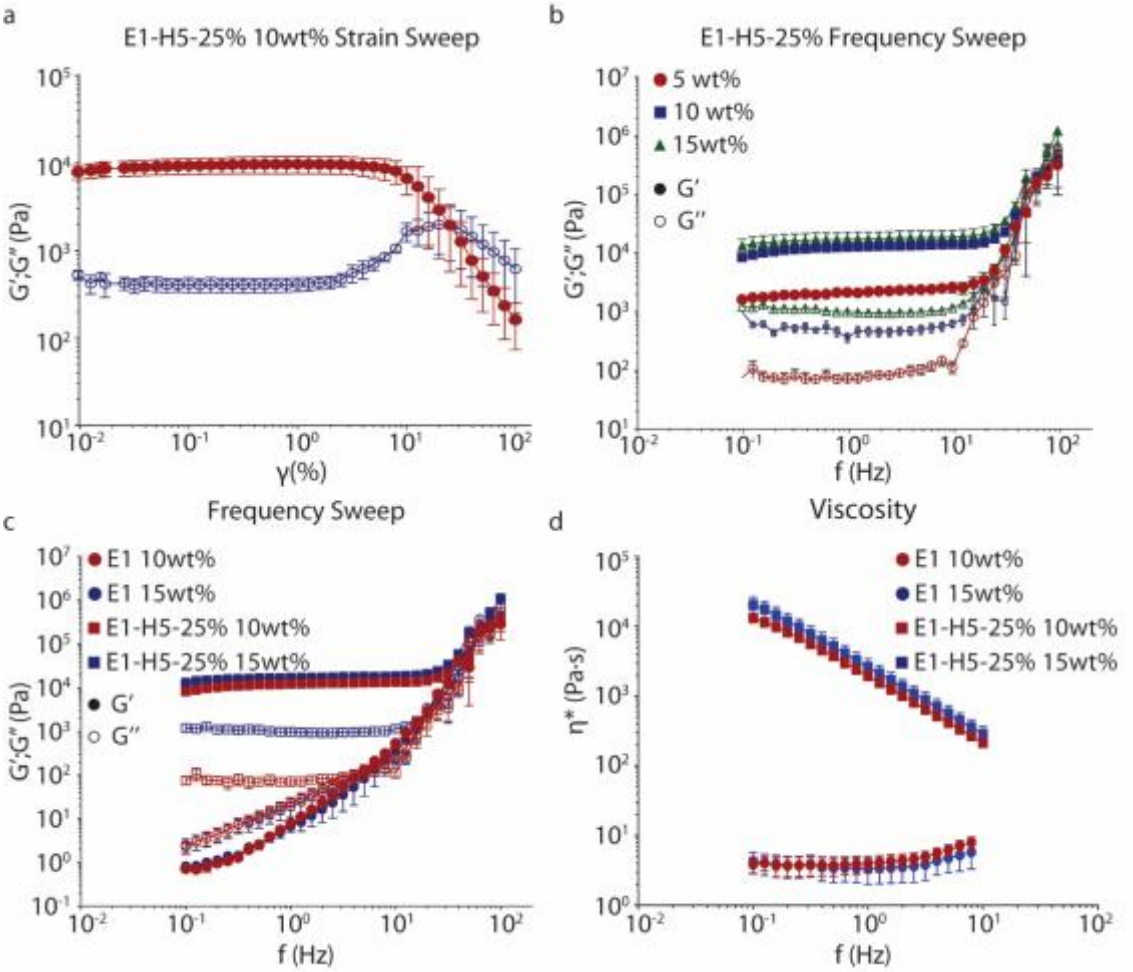


Figure 6.7: Rheology. (a) A strain sweep from 0.01 to 100% reveals the linear viscoelastic region (LVER) of POPs. (b) Frequency sweeps within the LVER (1% strain) reveal solid-like material properties for POPs which scale non-linearly with concentration. (c) ELPs show more liquid-like behavior (G">G') and decrease mechanical integrity compared to POPs. (d) POPs exhibit plastic, frequency dependent viscosity whereas ELPs behave as Newtonian fluids. All experiments in PBS after 30 min equilibration at 37°C.

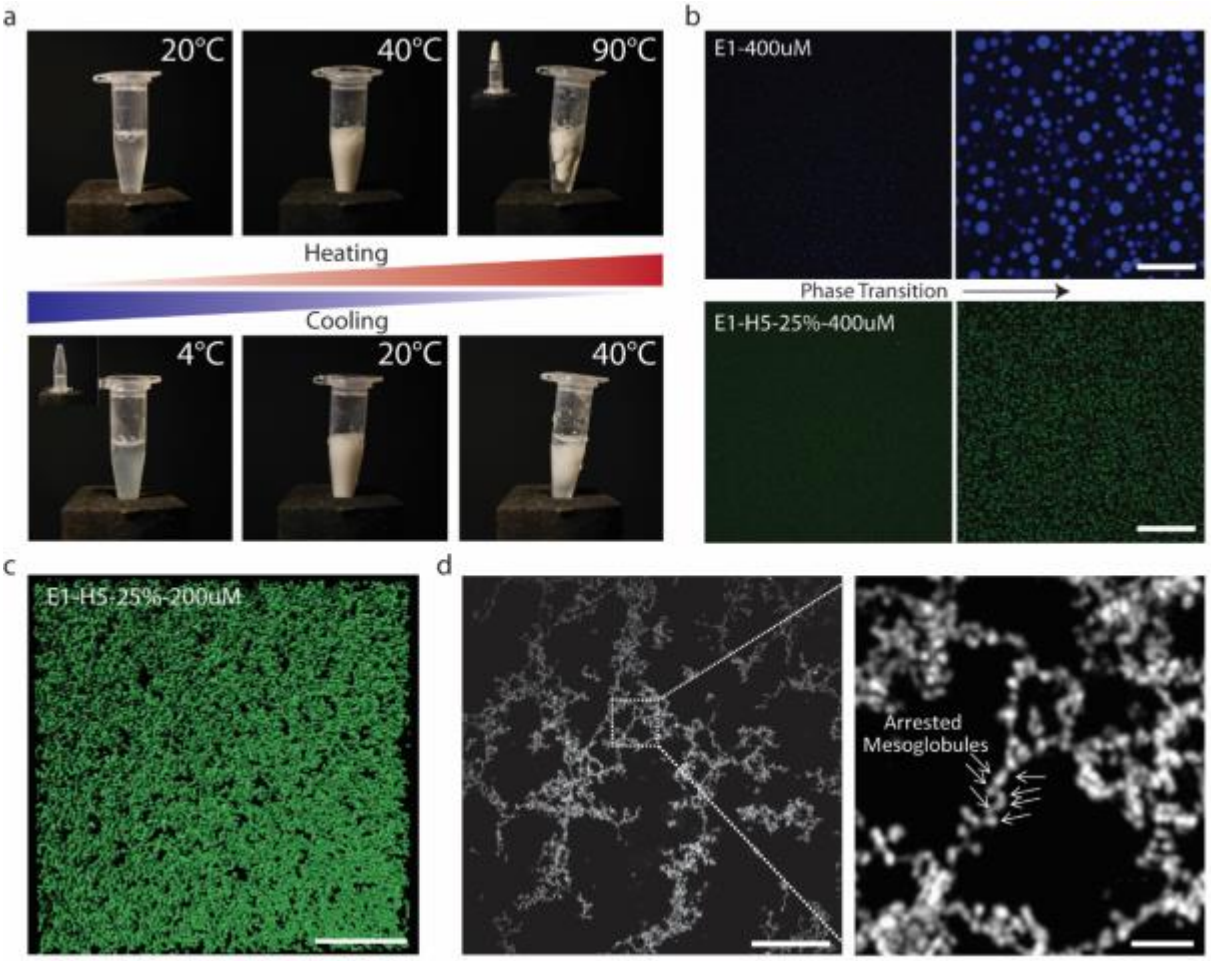


Figure 6.8: Arrested Phase Separation into Fractal Networks. (a) E1-H5-25% (2mM, PBS) aggregation during a heating and cooling cycle shows a reversible transition from an optically translucent liquid to an opaque solid-like structure (passes inversion test) with syneresis observed at higher temperatures. (b) At the microscale, E1 and E1-H5-25% (400uM, PBS) form liquid-like coacervates and fractal networks, respectively; scale bar 50um. (c) The intricacy of the network is more clearly seen with a 20um thick 3D reconstruction of E1-H5-25% (200uM, PBS); scale bar 50uM. (d) Network architecture at the meso scale is that of interconnected "beads on a string", as revealed by SIM; scale bars 10um (left) and 1um (right).

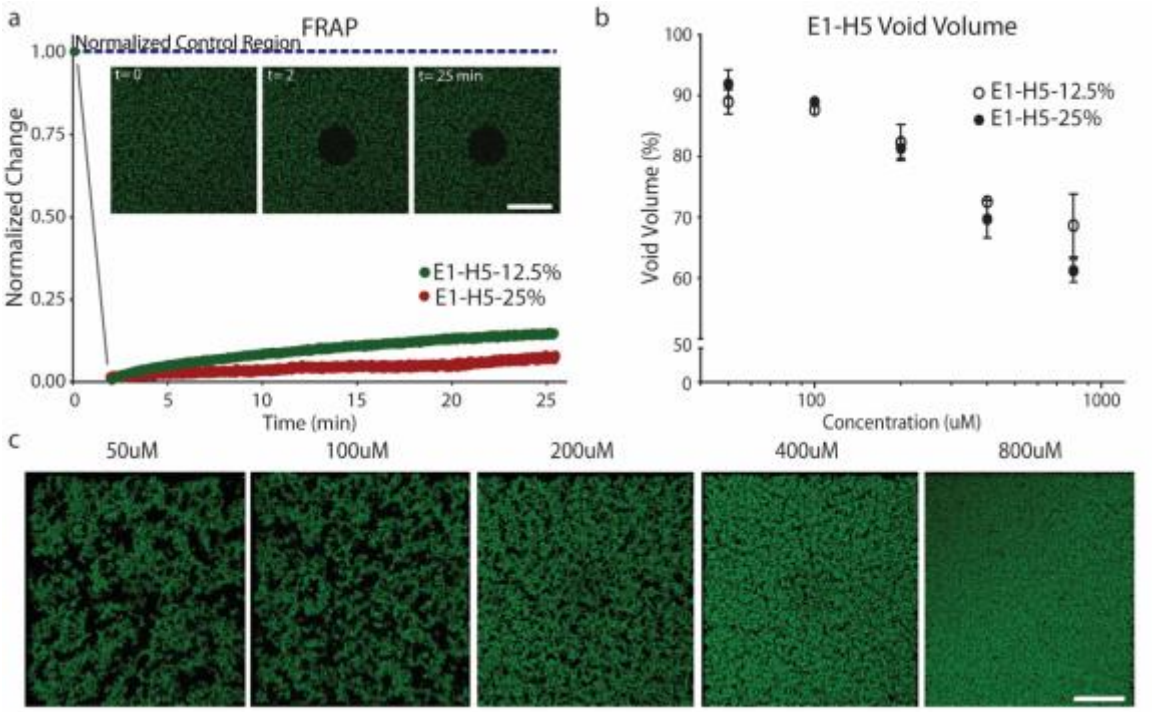


Figure 6.9: Network Stability and Void Volume. (a) As determined by the limited fluorescence recovery 25 min after bleaching, 12.5% and 25% networks have a high kinetic stability and limited liquid-like properties; Inset pictures are shown for E1-H5-25% at 400uM. (b-c) Void volumes can be tuned from 60-90% by altering polymer concentration. Scale bars are 50um.

6.7 *In Situ* Network Stability and Cell Penetration

For *in vivo* applications, thermo-responsive biopolymers possess inherent material advantages of biocompatibility and controllable *in situ* assembly. These advantages have been well exploited by disordered biopolymers as injectable depots[53-55]; however, their lack of porosity and mechanical stability have largely limited their potential for applications such as tissue engineering[11, 56]. Solid-like, porous, biopolymers networks which still exhibit this thermal assembly, therefore, have the potential alter the landscape of available applications.

To assess the *in vivo* capabilities of our POPs, we first injected them as subcutaneous depots to assess their pharmacokinetic (PK) properties. They were simultaneously compared to a fully disordered ELP of the same base sequence for comparison. Both polymers were labelled

with ^{125}I and injected to the subcutaneous space of the right flank. Blood samples were periodically taken to evaluate polymer release. Subcutaneous depots were also imaged using single-photon emission computed tomography (SPECT) to evaluate changes in depot volume and shape. POP depots showed significantly less polymer release (4.8% of initial dose) from the depot than their disordered counterparts (8.7% of initial dose) after 120 hours, despite their increased porosity and greater surface area (Fig. 6.10a). Importantly, terminal biodistribution analysis revealed no critical or unexpected accumulation in vital organs (Fig. 6.11). Visual inspection and SPECT analysis of the depots after injection reveals an even more striking difference between the partially ordered and disordered polymers (Fig. 6.10b-c). ELP injections expand in the *s.c.* space until they are not externally apparent. POP injections, on the other hand, retain their shape, forming large, palpable depots easily seen through the skin. Histological analysis of explanted POP depots after 120hrs also shows a significant penetrating cell mass at both the edges and the core of our depots along with collagen fiber formation, large blood vessel penetration, and capillary growth (Fig. 6.10d). Evaluation of ELP depots was not possible as they were too diffuse to be observable during dissection. Though further studies on the long-term effects of injected POPs are certainly warranted, their short-term behavior gives clear indication of the advantages arising from combinations of disordered behavior—stimuli-responsiveness—and ordered behavior—architecture and stability.

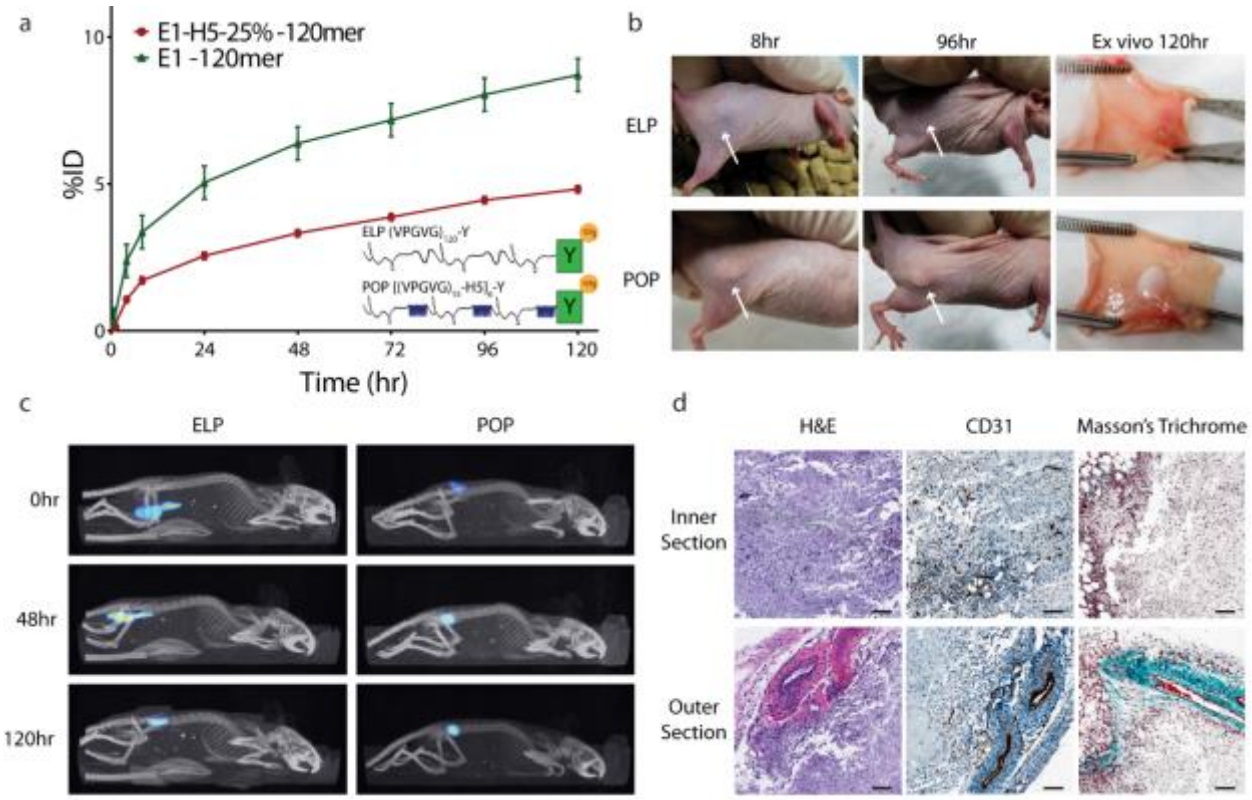


Figure 6.10: In Vivo Comparison of ELPs and POPs. (a) E1-H5-25% POP s.c. injections were significantly more stable than their E1 counterparts with just 5% of the injected dose degraded at 120hrs; 200ul 200uM injections; $p<0.05$ for all data points after 0hr. (b) Whereas ELPs diffuse into the s.c. space, POP depots were externally apparent, retaining the shape and volume of the initial injection up to dissection and ex vivo analysis. (c) Representative CT-SPECT images of the depots confirm increased diffusivity of ELPs and increased stability of POPs. (d) Histological analysis of POP depots reveals a high cell penetration with some collagen deposition and blood vessel and capillary formation; scale bar 100um.

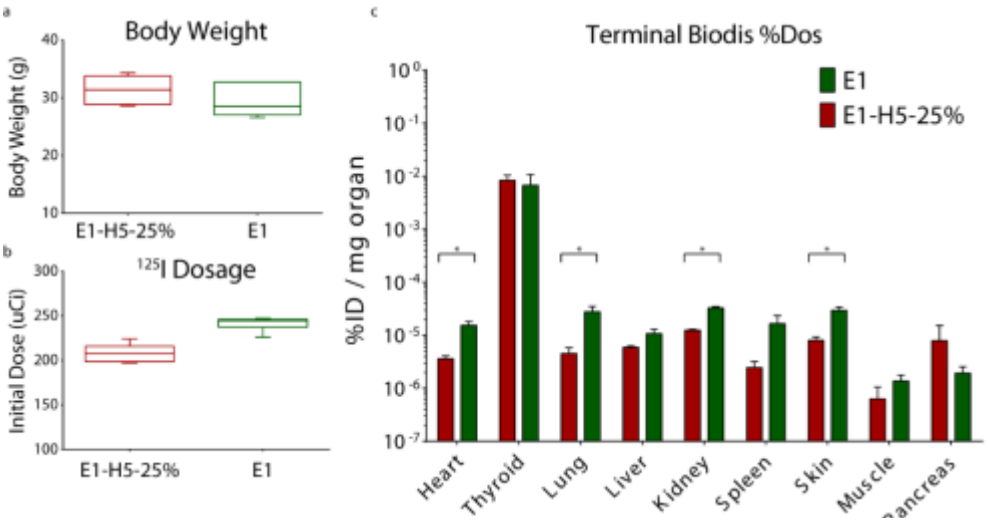


Figure 6.11: Biodistribution of POPs. (a) Body weight measures for all mice used are given along with the (b) dose of 125I for each group. Dosages were used for data normalization and differences in doses on reflect an increase in bound iodine for POPs which is not expected to be experimentally relevant. (c) Radiation measured for organs after 120hrs reveals some small distribution differences for POPs and ELPs, but none expected to be harmful.

6.8 Discussion

The interplay between ordered and disordered domains is a critical component for biological interactions. Using precisely designed recombinant proteins we have established a first of its kind platform to evaluate these interactions and their ability to produce exciting material properties. We have shown that properties of both disordered and ordered domains can be maintained in designer materials, while also combining to create emergent material properties unachievable using only the independent components. Specifically, our POPs retain the reversible, thermal-responsive phase behavior of disordered ELPs with these components driving phase separation and controlling the initial aggregation temperature. The structured, polyalanine domains alter the architecture of the aggregates and control the dissolution temperature. The resultant materials are fractal-like protein networks with highly tunable thermal hysteresis. Because the specific components of the polymers can be independently controlled, the properties of the materials can be orthogonally altered. With an eye towards applications in tissue engineering, we have also shown that these polymers assemble into 3D scaffolds *in vivo*, retain their shape, and are notably more stable than comparable disordered polymers. As the field of intrinsic protein disorder has expanded, knowledge of the biological importance of disorder-order interactions has also grown. Yet limited information exists on how these interactions may be manipulated or functionalized for biomedical application. Our biopolymer platform is a first and important step towards developing these design rules, and further study promises a new generation of functional protein materials.

6.9 Methods

Synthesis of polymer genes: All polymers were cloned into a modified pet24 vector using a previously described process known as recursive directional ligation by plasmid reconstruction (PRe-RDL)[4]. Briefly, single stranded oligomers encoding the desired sequences were annealed into cassettes with CC and GG overhangs. The overhangs enabled their concatemerization and ligation (Quick Ligase, NEB, Ipswich, MA) into the pet24 vector. Using this process, we created a library of elastin-like polypeptide and polyalanine cassettes which could be pieced together through multiple cycles of PRe-RDL to form the final partially ordered polymers. All of the base oligomer cassettes used for polymer construction can be found below. Plasmids were transfected into chemically competent Eb5α (EdgeBio, Gaithersburg, MD) cells for cloning and BL21(DE3) (EdgeBio, Gaithersburg, MD) cells for protein expression.

DNA Cassettes for Pre-RDL:

E1	Forward	TGTGGGTGTTCCGGGCGTAGGTGTCCCAGGTGTGGGCGTACCGGGCGTTGGTGTTCCTGGTGTCGGCGTGCCGGG
	Reverse	CGGCACGCCGACACCAGGAACACCAACGCCCGGTACGCCCACACCTGGGACACCTACGCCCGGAACACCCACACC
E2	Forward	CGTGGGTGTTCCGGGCGTAGGTGTCCCAGGTGCGGGCGTACCGGGCGTTGGTGTTCCTGGTGTCGGCGTGCCGGG
	Reverse	CGGCACGCCGACACCAGGAACACCAACGCCCGGTACGCCCGCACCTGGGACACCTACGCCCGGAACACCCACGCC
E3	Forward	CGCCGGAGTGCCAGGCGTGGGTGTTCCAGGAGCAGGCGTTCCAGGTGTGGGTGTTCCTGG
	Reverse	AGGAACACCCACACCTGGAACGCCTGCTCCTGGAACACCCACGCCTGGC

		ACTCCGGCGCC
H1	Forward	TGCGGCCGCAGCTGCGGCGGCAGCCGCGGCTGCCGCGGCTGCAGCGGCAGCCGCGGCTGCGGCGGCCGCAGCTGCGGG
	Reverse	CGCAGCTGCGGCCGCCGCAGCCGCGGCTGCCGCTGCAGCCGCGGCAGCCGCGGCTGCCGCCGCAGCTGCGGCCGCACC
H2	Forward	TAAAGCGGCCGCAGCTGCGGCGGCAGCCGCGGCTGCCGCGGCTGCAGCGGCAGCCGCGGCTGCGGCGGCCGCAGCTGCGAAAGG
	Reverse	TTTCGCAGCTGCGGCCGCCGCAGCCGCGGCTGCCGCTGCAGCCGCGGCAGCCGCGGCTGCCGCCGCAGCTGCGGCCGCTTTACC
H3	Forward	TAAAGCGGCCGCAGCTAAAGCCGCGGCAGCGAAAGCAGCCGCGGCGAAAGCCGCAGCTGCGAAAGCGGCAGCCGCGAAGGG
	Reverse	CTTCGCGGCTGCCGCTTTCGCAGCTGCGGCTTTCGCCGCGGCTGCTTTCGCTGCCGCGGCTTTAGCTGCGGCCGCTTTACC
H5	Forward	TGATGCGGCCGCAGCTGCGGCGGCAGCCGCGGCTGCCGCGGCTGCAGCGGCAGCCGCGGCTGCGGCGGCCGCAGCTGCGAAAGG
	Reverse	TTTCGCAGCTGCGGCCGCCGCAGCCGCGGCTGCCGCTGCAGCCGCGGCAGCCGCGGCTGCCGCCGCAGCTGCGGCCGCATCACC

Expression and purification of POPs: For protein expression, 5mL starter cultures were grown overnight from -80°C DMSO stocks. Cells were then pelleted, resuspended in 1mL of terrific broth, and used, along with 1mL 100µg mL^{-1} of kanamycin (EMD Millipore, Billerica, MA) to inoculate 1L of media. Cells were shaken at 200rpm for 8hrs at 25°C before induction. For induction of protein expression, 1mL of 1M isopropyl β-D-1-thiogalactopyranoside (Goldbio, St. Louis, MO) was added to the flask and cultures were placed at 16°C and 200 rpm overnight. Expression at lower temperature was necessary to prevent the formation of truncation products at

ELP-polyalanine junctions. Cells were then pelleted and resuspended in 10mL of 1X PBS for every 1L of culture grown. Pulse sonication on ice, with a total active time of 3 minutes, was used to lyse cells. Cell lysates were treated with 10% PEI (MP Biomedical, Santa Ana, CA) (2ml L^{-1} culture) to remove contaminating DNA and centrifuged at 14k rpm for 10min at 4˚C. Polymer was purified from the resulting soluble fraction using a modified version of inverse thermal cycling[9]. The fraction was heated to 65˚C or until a phase separation was observed. For more hydrophilic polymers, this often required the addition of 1-2M NaCl to depress the transition temperature. Once aggregated, the polymer solutions were centrifuged at 14k rpm for 10min at 35˚C, and the resulting pellet was re-suspended in 5-10ml PBS. The heating and cooling centrifugation cycles were repeated 2-3 more times until a purity of 95% was achieved, as analyzed by SDS-PAGE. Pure polymers were dialyzed at 4°C with frequent water changes for 2 days and lyophilized for storage.

Secondary structure characterization: Circular dichroism experiments were carried using an Aviv Model 202 instrument and 1mm quartz cells (Hellma USA, Plainview, NY). Unless otherwise noted, scans were carried out in PBS (pH=7.4) with a polymer concentration of 10uM. Polymers were scanned in triplicate from 260nm to 185nm in 1nm steps with a 1s averaging time. Data points with a dynode voltage above 500V were ignored for analysis. All measurements were done at 20°C unless otherwise specified. Temperature ramping was done in 5°C/min increments with a 1 min equilibration at each step.

For NMR, polymers were grown in M9 minimal media with ^{15}N-NH_4Cl and ^{13}C-Glucose (Cambridge Isotopes, Tewksbury, MA) as the only nitrogen and carbon sources to ensure protein labelling. Samples were prepared in PBS (pH=7.4) unless otherwise noted. All NMR spectra were collected on a INOVA 600 (Varian Instruments, Palo Alto, CA) spectrometer with a triple

resonance cryoprobe equipped with a z-field gradient coil. Resonance assignments were made using a set of triple resonance experiments including HNCO, HN(CA)CO, HN(CO)CA, HNCA, HCAN, and HCA(CO)N. The NMR spectra were processed using NMRpipe[57], and were analyzed using NMRviewJ. Chemical shifts in the proton dimension were referenced relative to TMSP (trimethylsilylpropanoic acid) as 0 ppm. Quantification of helicity was accomplished using the identified alanine peaks of the H(N)CO spectra for E1-H2-25%. Chemical shift positions were placed on a spectrum of values ranging from fully disordered (177.19 ppm) to fully helical (180.78 ppm), as determined Vendruscolo *et al*.[58, 59] and the central alanine peak of the 15°C H(N)CO respectively, producing the values in Sup. Table 2. The method to calculate helicity was adapted from δ2D algorithm developed by Vendruscolo *et al*.[59] Alanines corresponding to carbon chemical shifts of peaks 2-7 were designated as fringe amino acids at the edges of the helix. This designation is consistent with our helix-coil transition theory prediction in which 6 alanines occur at values lower than the core set. All other alanines were assumed to be in the helix core. A subsequent averaging of the helicity values produces a helicity for each H2 polyalanine domain of 91%.

Temperature-dependent turbidimetry: The transition temperature (Tt) of each sample was determined by monitoring the optical density at 350nm as a function of temperature on a UV-vis spectrophotometer (Cary 300 Bio; Varian Instruments, Palo Alto, CA) equipped with a multicell thermoelectric temperature controller. The Tt was defined as the point of greatest inflection (maximum of the first derivative) for the optical density. Unless otherwise stated, all samples were heated and cooled at 1 °C min^{-1} in PBS at concentrations between 10 and 1000um.

Molecular dynamics simulations: The phenomenological simulations were designed to test the role of having two energy scales on the coarse structural features. We chose the interaction strengths of the ELP beads such that this range would span from highly soluble to aggregating polymers. This was

quantified by running simulations with a range of energies and after equilibrating for 100ns, decreasing these interaction strengths by .05 kcal/mol every 25ns. We then quantified the number of polymers in the largest cluster, where two proteins were considered interacting if two beads were within 8 Angstroms, as a proxy for aggregation versus solubility in our simulations. As shown in figure xxx, these polymers were strongly aggregating with an interaction strength of 0.35 kcal/mol and readily disaggregated when that interaction dropped to 0.25 kcal/mol. As such, we used a range of interactions strengths for the ELPs that spanned at least 0.05 kcal/mol to 0.40 kcal/mol. This range of interaction strengths is our simulation equivalent to increasing the temperature of the system from below the LCST to above the LCST. Unfortunately, without any further constraints, we cannot be more quantitative in the scaling between the strength of our interactions and the experimental equivalent temperatures. We used a similar technique to parameterize the alanine domain bead interaction strength shown in figure xxx+1. Here our constraint in choosing an interaction strength is based on being strong enough to push it significantly into the aggregation prone regime. As such, we used interaction strengths of at least 1kcal/mol for the alanine beads.

To test for effects related to hysteresis we utilized two different schemes for initial conditions. The first scheme, denoted the dimer initial conditions, was designed to create states that we think are representative of the pathway that the system will pass through as it approaches the LCST from below. Simulations of two proteins were equilibrated for xxxns in a simulation box of xxxA. This allowed the alanine domains in these dimers to pre-aggregate into a core. 25 different conformations of these dimers were then randomly placed in the simulations for the full system. At high ELP interaction strengths these simulations docked together. This means that the ELPs that are exposed around the alanine cores find each other. There is some degree of alanine cores merging together into larger cores that converge toward their thermodynamically favorable radius.

The second scheme, denoted the coil initial conditions, was designed to create a thermodynamically equilibrated aggregated state that we think the dimer initial condition simulations

would eventually converge toward. We started the simulation with each polymer generated randomly. The only correction was to prevent steric clashes. These simulations showed a rapid initial collapse as the alanine domains found other alanine domains, and, if the ELP domains were above the LCST, the collapse of ELP domains as well. These simulations converged toward conformations with clusters of alanine domains that were well connected. After equilibrating for 100ns, the interaction strength of the ELPs were decreased by 0.10 kcal/mol every 25ns to model crossing from above the LCST to below the LCST. These simulations showed swelling as the ELPs no longer favored being in a high density but the connectivity of alanine domains between the two domains prevented the system from separating.

Fluorescence imaging and analysis: POPs were fluorescently labeled using Alexa Fluor 488 NHS Ester (Thermo Fisher, Waltham, MA) with a reaction efficiency of 20%. Excess dye was removed with dialysis and polymers were lyophilized for storage. For all experiments, the dyed polymers were diluted into an undyed stock such that no more than 5% of POPs in solution were labelled. Confocal images were taken on a Zeiss 710 inverted microscope with temperature controlled incubation. To prevent dehydration, 50ul of sample solution was added to 384 well #1.5 glass bottom plates (Cellvis, Mountain View, CA) for imaging. Solutions were added below the Tt and allowed to transition and equilibrated for 5 minutes on the microscope stage. For FRAP experiments, samples were equilibrated for 30 min to prevent thermal movement of the focusing stage, and fluorescence intensity analysis was done using Zen software (ZEISS Microscopy, Jena, Germany). For void volume analysis, 20um image stacks were taken with a pinhole size of 1 Airy unit and vertical slice intervals of 230 nm. Three dimensional reconstructions of the resultant networks and quantification of their void volume was done in IMARIS 8 (Bitplane. Belfast, Ireland). Surface renders were constructed with a minimum object detail of 200nm and local background thresholding with the diameter of the largest sphere that fits into the object set a 1um. A consistent minimum threshold of 1000 FU was used across

samples. Network fractal dimensions were determined using the 2D box counting algorithm from the FracLac plugin for ImagJ [60, 61].

Pharmacokinetic and SPECT analysis: All constructs were prepared at 500μM in sterilized PBS and reacted with 125Iodine (Perkin Elmer, Boston MA) in Pierce® pre-coated IODOGEN tubes (Fisher Scientific, Hampton, NH)[62]. The product was centrifugally purified through 40K MWCO Zeba Spin Desalting Columns (Thermo Scientific, Rockford, IL) at 2500 rpm for 3 min at 4°C to remove unreacted radioiodine from the conjugate. After labeling, each construct was diluted down to a final biopolymer concentration of 250μM. The resulting activity dose for the helical construct was 1.18 mCi mL^{-1}, while the unimer construct dose was 1.37 mCi mL^{-1}.

50uL of the POP was prepared in an Eppendorf tube at 63μCi to provide a reference imaging standard. Prior to either the depot injection, blood draw, or single-photon emission computed tomography (SPECT) imaging, each mouse was anesthetized using a 1.6% isoflurane vaporizer feed at an O2 flow rate of 0.6 L min^{-1}. For depot injections, each mouse received a soluble 200uL injection of their respective solution at 250μM into the subcutaneous space on the right hind flank. The whole body activity of the mouse was then measured in an AtomLab 400 dose calibrator (Biodex, Shirley, NY). A total of 12 athymic nude mice (6 per group) were used for pharmacokinetic analysis of depot stability and distribution. An initial 10μL blood sample was drawn and pipetted into 1000mg mL^{-1} heparin with subsequent blood draws at time points of 45min, 4h, 8h, 24h, 48h, 72h, 96, and 120h to determine the release profile for the depots. 6 total athylmic nude mice also were imaged using SPECT at time points of 0, 48, and 120 hrs.

Mice were then transferred under anesthesia to the bed of the U-SPECT-II/CT for imaging using a 0.350 collimator (MILabs B.V., Utrecht, Netherlands) courtesy of G. Al Johnson in the Duke CIVM. Anesthesia was maintained with a 1.6% isoflurane feed at an O2 flow rate of 0.6 L min-

[1]. SPECT acquisition was conducted over a time frame of 15 minutes in 'list-mode' and at a 'fine' step-mode. Upon completion, a subsequent CT scan was carried out at a current of 615μA and a voltage of 65kV. Mice were then returned to their cages. Post-imaging SPECT reconstruction was carried out using MILabs proprietary software without decay correction and centered on the ^{125}I photon range of 15-45 keV. All images were reconstructed at a voxel size of 0.2 mm. Reconstructed SPECT images were then registered with their corresponding CT scans to provide spatial alignment for anatomical reference.

Upon completion of the study, all mice were euthanized and dissected. The subcutaneous depots were excised and visually examined for physical differences. In addition, the heart, thyroid, lungs, liver, kidneys, spleen, skin, muscle and pancreas were collected and analyzed using a Wallac 1282 Gamma Counter (Perkin Elmer, Boston, MA) to determine the relative biodistribution of the different constructs. All blood samples and the set of pk standards were similarly analyzed using the gamma counter. The counts per minute detected for each sample were converted to their corresponding activity. Blood samples were then scaled to determine the total amount in circulation according to the formula Total = CPM/.01*BW*72ml/kg[63].

Depot retention was analyzed by measuring the total photon intensity of the depot SPECT image in ImageJ (NIH, public domain). Measured photon intensity was converted to total depot activity using a calibration factor determined from the imaging standard. This calibration was determined by performing a linear regression of the known activities of the standard over time against the corresponding SPECT intensity measurements. The factor was applied to each depot and the calculated activity compared against the original whole body injected dose at 0h to determine its percent retention.

6.10 Acknowledgments

We are grateful to Kiersten Ruff for stimulating discussions. Grants from the National Science Foundation (MCB1614766 to RVP, DMR 11-21107 to AC) and the National Institutes of Health (RO1- GM061232 to AC) supported this work. TSH is a graduate student scholar of the Center for Biological Systems Engineering at Washington University in St. Louis. Additionally we appreciate the use of the Duke SMIF core facility and the Duck Light Microscopy Core Facility (LMCF).

6.11 References

1. Keten, S., et al., *Nanoconfinement controls stiffness, strength and mechanical toughness of beta-sheet crystals in silk.* Nat Mater, 2010. **9**(4): p. 359-67.
2. van der Lee, R., et al., *Classification of intrinsically disordered regions and proteins.* Chem Rev, 2014. **114**(13): p. 6589-631.
3. Van Roey, K., et al., *Short linear motifs: ubiquitous and functionally diverse protein interaction modules directing cell regulation.* Chem Rev, 2014. **114**(13): p. 6733-78.
4. McDaniel, J.R., et al., *Recursive directional ligation by plasmid reconstruction allows rapid and seamless cloning of oligomeric genes.* Biomacromolecules, 2010. **11**(4): p. 944-52.
5. McDaniel, J.R., D.C. Radford, and A. Chilkoti, *A unified model for de novo design of elastin-like polypeptides with tunable inverse transition temperatures.* Biomacromolecules, 2013. **14**(8): p. 2866-72.
6. Li, N.K., et al., *Molecular description of the LCST behavior of an elastin-like polypeptide.* Biomacromolecules, 2014. **15**(10): p. 3522-30.

7. Meyer, D.E. and A. Chilkoti, *Quantification of the effects of chain length and concentration on the thermal behavior of elastin-like polypeptides.* Biomacromolecules, 2004. **5**(3): p. 846-51.

8. Roberts, S., M. Dzuricky, and A. Chilkoti, *Elastin-like polypeptides as models of intrinsically disordered proteins.* FEBS Lett, 2015. **589**(19 Pt A): p. 2477-86.

9. Meyer, D.E. and A. Chilkoti, *Purification of recombinant proteins by fusion with thermally-responsive polypeptides.* Nat Biotechnol, 1999. **17**(11): p. 1112-5.

10. McDaniel, J.R., D.J. Callahan, and A. Chilkoti, *Drug delivery to solid tumors by elastin-like polypeptides.* Adv Drug Deliv Rev, 2010. **62**(15): p. 1456-67.

11. Nettles, D.L., A. Chilkoti, and L.A. Setton, *Applications of elastin-like polypeptides in tissue engineering.* Adv Drug Deliv Rev, 2010. **62**(15): p. 1479-85.

12. Pometun, M.S., E.Y. Chekmenev, and R.J. Wittebort, *Quantitative observation of backbone disorder in native elastin.* J Biol Chem, 2004. **279**(9): p. 7982-7.

13. Chakrabartty, A. and R. Baldwin, *Stability of α-Helices.* 1995. **46**: p. 141-176.

14. Farmer, R.S. and K.L. Kiick, *Conformational behavior of chemically reactive alanine-rich repetitive protein polymers.* Biomacromolecules, 2005. **6**(3): p. 1531-9.

15. Bernacki, J.P. and R.M. Murphy, *Length-dependent aggregation of uninterrupted polyalanine peptides.* Biochemistry, 2011. **50**(43): p. 9200-11.

16. Miller, J.S., R.J. Kennedy, and D.S. Kemp, *Solubilized, Spaced Polyalanines: A Context-Free System for Determining Amino Acid α-Helix Propensities.* Journal of the American Chemical Society, 2002. **124**(6): p. 945-962.

17. Gosline, J., et al., *Elastic proteins: biological roles and mechanical properties.* Philos Trans R Soc Lond B Biol Sci, 2002. **357**(1418): p. 121-32.

18. Lillie, M.A. and J.M. Gosline, *The viscoelastic basis for the tensile strength of elastin.* Int J Biol Macromol, 2002. **30**(2): p. 119-27.

19. Li, D.Y., et al., *Elastin is an essential determinant of arterial morphogenesis.* Nature, 1998. **393**(6682): p. 276-80.

20. Shadwick, R.E., *Mechanical design in arteries.* J Exp Biol, 1999. **202**(Pt 23): p. 3305-13.

21. Tamburro, A.M., A. Pepe, and B. Bochicchio, *Localizing alpha-helices in human tropoelastin: assembly of the elastin "puzzle".* Biochemistry, 2006. **45**(31): p. 9518-30.

22. Tamburro, A.M., B. Bochicchio, and A. Pepe, *Dissection of human tropoelastin: exon-by-exon chemical synthesis and related conformational studies.* Biochemistry, 2003. **42**(45): p. 13347-62.

23. Miao, M., et al., *Structural determinants of cross-linking and hydrophobic domains for self-assembly of elastin-like polypeptides.* Biochemistry, 2005. **44**(43): p. 14367-75.

24. Miao, M., et al., *Sequence and domain arrangements influence mechanical properties of elastin-like polymeric elastomers.* Biopolymers, 2013. **99**(6): p. 392-407.

25. Vrhovski, B. and A.S. Weiss, *Biochemistry of tropoelastin.* Eur J Biochem, 1998. **258**(1): p. 1-18.

26. Micsonai, A., et al., *Accurate secondary structure prediction and fold recognition for circular dichroism spectroscopy.* Proc Natl Acad Sci U S A, 2015. **112**(24): p. E3095-103.

27. Bochicchio, B., A. Pepe, and A.M. Tamburro, *Investigating by CD the molecular mechanism of elasticity of elastomeric proteins.* Chirality, 2008. **20**(9): p. 985-94.

28. Muñoz, V. and L. Serrano, *Elucidating the folding problem of helical peptides using empirical parameters.* Nature Structural Biology, 1994. **1**(6): p. 399-409.

29. Muñoz, V. and L. Serrano, *Elucidating the Folding Problem of Helical Peptides using Empirical Parameters. II†. Helix Macrodipole Effects and Rational Modification of the Helical Content of Natural Peptides.* Journal of Molecular Biology, 1995. **245**(3): p. 275-296.

30. Munoz, V. and L. Serrano, *Elucidating the folding problem of helical peptides using empirical parameters. III. Temperature and pH dependence.* J Mol Biol, 1995. **245**(3): p. 297-308.

31. Lacroix, E., A.R. Viguera, and L. Serrano, *Elucidating the folding problem of alpha-helices: local motifs, long-range electrostatics, ionic-strength dependence and prediction of NMR parameters.* J Mol Biol, 1998. **284**(1): p. 173-91.

32. Wright, E.R. and V.P. Conticello, *Self-assembly of block copolymers derived from elastin-mimetic polypeptide sequences.* 2002. **54**: p. 1057-1073.

33. Wright, E.R., et al., *Thermoplastic elastomer hydrogels via self-assembly of an elastin-mimetic triblock polypeptide.* Advanced Functional Materials, 2002. **12**(2): p. 149-154.

34. Glassman, M.J., et al., *Toughening of Thermoresponsive Arrested Networks of Elastin-Like Polypeptides To Engineer Cytocompatible Tissue Scaffolds.* Biomacromolecules, 2016. **17**(2): p. 415-26.

35. Kim, W. and E.L. Chaikof, *Recombinant elastin-mimetic biomaterials: Emerging applications in medicine.* Adv Drug Deliv Rev, 2010. **62**(15): p. 1468-78.

36. Herrero-Vanrell, R., et al., *Self-assembled particles of an elastin-like polymer as vehicles for controlled drug release.* J Control Release, 2005. **102**(1): p. 113-22.

37. Reguera, J., et al., *Thermal Behavior and Kinetic Analysis of the Chain Unfolding and Refolding and of the Concomitant Nonpolar Solvation and Desolvation of Two Elastin-like Polymers.* Macromolecules, 2003. **36**(22): p. 8470-8476.

38. Cho, Y., et al., *Hydrogen bonding of beta-turn structure is stabilized in D(2)O.* J Am Chem Soc, 2009. **131**(42): p. 15188-93.

39. Ding, F., et al., *Mechanism for the alpha-helix to beta-hairpin transition.* Proteins, 2003. **53**(2): p. 220-8.

40. Urry, D.W. and T.H. Ji, *Distortions in circular dichroism patterns of particulate (or membranous) systems.* Arch Biochem Biophys, 1968. **128**(3): p. 802-7.

41. Urry, D.W., B. Starcher, and S.M. Partridge, *Coacervation of solubilized elastin effects a notable conformational change.* Nature, 1969. **222**(5195): p. 795-6.

42. Urry, D.W., T.A. Hinners, and L. Masotti, *Calculation of distorted circular dichroism curves for poly-L-glutamic acid suspensions.* Arch Biochem Biophys, 1970. **137**(1): p. 214-21.

43. Urry, D.W. and J. Krivacic, *Differential scatter of left and right circularly polarized light by optically active particulate systems.* Proc Natl Acad Sci U S A, 1970. **65**(4): p. 845-52.

44. Sagle, L.B., et al., *Investigating the hydrogen-bonding model of urea denaturation.* J Am Chem Soc, 2009. **131**(26): p. 9304-10.

45. Muiznieks, L.D., S.A. Jensen, and A.S. Weiss, *Structural changes and facilitated association of tropoelastin.* Archives of Biochemistry and Biophysics, 2003. **410**(2): p. 317-323.

46. Yeo, G.C., F.W. Keeley, and A.S. Weiss, *Coacervation of tropoelastin.* Adv Colloid Interface Sci, 2011. **167**(1-2): p. 94-103.

47. Cirulis, J.T., F.W. Keeley, and D.F. James, *Viscoelastic properties and gelation of an elastin-like polypeptide.* Journal of Rheology, 2009. **53**(5): p. 1215.

48. Chen, D.T.N., et al., *Rheology of Soft Materials.* Annual Review of Condensed Matter Physics, 2010. **1**(1): p. 301-322.

49. Gustafsson, M.G.L., *Surpassing the lateral resolution limit by a factor of two using structured illumination microscopy. SHORT COMMUNICATION.* Journal of Microscopy, 2000. **198**(2): p. 82-87.

50. Gustafsson, M.G., *Nonlinear structured-illumination microscopy: wide-field fluorescence imaging with theoretically unlimited resolution.* Proc Natl Acad Sci U S A, 2005. **102**(37): p. 13081-6.

51. Tu, Y., S.G. Wise, and A.S. Weiss, *Stages in tropoelastin coalescence during synthetic elastin hydrogel formation.* Micron, 2010. **41**(3): p. 268-72.

52. Clarke, A.W., et al., *Tropoelastin massively associates during coacervation to form quantized protein spheres.* Biochemistry, 2006. **45**(33): p. 9989-96.

53. Adams, S.B., Jr., et al., *Sustained release of antibiotics from injectable and thermally responsive polypeptide depots.* J Biomed Mater Res B Appl Biomater, 2009. **90**(1): p. 67-74.

54. Amiram, M., et al., *Injectable protease-operated depots of glucagon-like peptide-1 provide extended and tunable glucose control.* Proc Natl Acad Sci U S A, 2013. **110**(8): p. 2792-7.

55. MacEwan, S.R. and A. Chilkoti, *Elastin-like polypeptides: biomedical applications of tunable biopolymers.* Biopolymers, 2010. **94**(1): p. 60-77.

56. Nettles, D.L., et al., *In situ crosslinking elastin-like polypeptide gels for application to articular cartilage repair in a goat osteochondral defect model.* Tissue Eng Part A, 2008. **14**(7): p. 1133-40.

57. Delaglio, F., et al., *NMRPipe: A multidimensional spectral processing system based on UNIX pipes.* Journal of Biomolecular NMR, 1995. **6**(3).

58. De Simone, A., et al., *Accurate random coil chemical shifts from an analysis of loop regions in native states of proteins.* J Am Chem Soc, 2009. **131**(45): p. 16332-3.

59. Camilloni, C., et al., *Determination of secondary structure populations in disordered states of proteins using nuclear magnetic resonance chemical shifts.* Biochemistry, 2012. **51**(11): p. 2224-31.

60. Schneider, C.A., W.S. Rasband, and K.W. Eliceiri, *NIH Image to ImageJ: 25 years of image analysis.* Nature Methods, 2012. **9**(7): p. 671-675.

61. A., K., *Karperien, A. FracLac for Image J, version 2.5*. 1999–2012: http://rsb.info.nih.gov/ij/plugins/fraclac/FLHelp/Introduction.htm.

62. Wood, W.G., C. Wachter, and P.C. Scriba, *Experiences Using Chloramine-T and 1,3,4,6-Tetrachloro-3-Alpha,6-Alpha-Diphenylglycoluril (Iodogen) for Radioiodination of Materials for Radioimmunoassay.* Journal of Clinical Chemistry and Clinical Biochemistry, 1981. **19**(10): p. 1051-1056.

63. Diehl, K.H., et al., *A good practice guide to the administration of substances and removal of blood, including routes and volumes.* J Appl Toxicol, 2001. **21**(1): p. 15-23.

Chapter 7

Conclusion

This thesis has brushed on a handful of different topics that all relate to cellular regulation ranging from protein binding to questions regarding phase transitions. Most of this work has left more questions unanswered than have been answered but we hope that this work opens the door for new directions in these projects.

Our work on coupled folding and binding in chapter 2 describes the methodology to redesign disordered proteins to titrate specific properties while holding other properties we expect to be important fixed. This work is intended not to be a standalone result but part of a larger strategy for understanding the role of disorder in coupled folding and binding mechanisms and how those mechanisms could be important to cells. We have been working in collaboration with Jane Clarke's lab to design experiments that can more directly probe these types of questions.

Chapter 3's work on the E4K4 peptide makes an important prediction for the field that the structure of the helical conformations might be the result of dramatic pKa shifts instead of the widely-believed salt bridge model. This model is supported by preliminary experimental data in collaboration with Ammon Posey which is still ongoing. This work illustrates the need for robust calculators for pKa values that works robustly for disordered proteins. Our implementation of GADIS for this endeavor is not transferable to systems where there isn't an already observed strong contradiction between the experimentally observed ensemble and the ensemble observed in simulations. If pKa shifts are more common than the field expects, we expect that the majority will not be associated with these needed strong contradictions. As such,

we have begun the initial work on a pKa calculator that we can couple to ABSINTH simulations in a Markov State Model.

The disordered linkers between domains plays a large role in controlling the driving forces for phase separation as well as the possibility of pushing the cell into a gelated state as we showed in chapter 4. This work has an abstract lattice to real space conversion of each lattice site being somewhere around 7 residues. In further work we hope to pushed towards more explicit predictions on the magnitude of the excluded volume of disordered regions. This should help narrow down what range of excluded volumes that are relevant for biology to explore. We are currently in discussions about designing a library of linkers with excluded volumes that span a large range. The plan is to quantify the excluded volume and end-to-end length through all atom simulations and measure the corresponding critical concentrations experimentally. We hope that this will ground us on the phase diagram and inform us if the gelation state is physically realizable.

In chapter 5 we showed a phase separated droplet observed in cells can have a well defined sub-organization inside nucleolus, a phase separated droplet in the cell nucleus. The Brangwynne lab reconstituted this behavior with the principle cellular components in vitro and we built a coarse grained model of how we expect the different proteins to interact which reproduces this behavior. Our work illustrated how different types of interacts contribute to the different surface tensions that yield the observed organization. Examining a single phase separated droplet merely scratches the surface though. There are many different types of phase separated droplets in cells and we hope that this work promotes the idea enough that other labs look for this type of behavior. One such example of this is an ongoing collaboration with Jingyi

Fei examining nuclear speckles. Similarly, we appear to be seeing spatial organization inside the droplet, but not to the extreme that is seen in the nucleolus.

Finally, in chapter 6, we examine a synthetic system that was designed by Stefan Roberts in Ashutosh Chilkoti's lab. This system was designed to have a strong and predictable hysteresis for use in medical applications. We built a coarse grained model of these polymers in order to study why these polymers have such a strong hysteresis. These polymers were inspired by tropoelastin which, on a coarse level, has a similar architecture: proline, valine, and glycine rich stretches mixed with alanine rich stretches. Our model of hysteresis for this synthetic system could have strong parallels with how elastin is able to be transported stably and then form a robust plastic material. Additionally, this analysis points to a possible pitfall that cells probably avoid in designing proteins that phase separate: the problem of mixing two distinct energy scales in a single protein.